The Dissertation Committee for Austin Williamson Reynolds certifies that this is the approved version of the following Dissertation:

Investigating Regional Human Population Histories in North America Using Genomics

Committee:

Deborah Bolnick, Supervisor

Daniel Bolnick

Mark Kirkpatrick

Patricia Galloway

Rasmus Nielsen

Investigating Regional Human Population Histories in North America Using Genomics

by

Austin Williamson Reynolds

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August 2018

Dedication

To my wife, for putting up with me.

Acknowledgements

As this will likely be the only accomplishment of note in my life, it is necessary not only to acknowledge those directly involved in my dissertation's completion, but also those who got me here in the first place. My primary mentors and role models have been smart, independent, kind, generous, and defiant women. Be they teachers, supervisors, family, or friends, they have been instrumental in shaping my life. I owe them a debt I can never hope to repay.

To my wife, your constant support throughout this whole process is the only reason I have made it here. Thanks to my family, particularly my parents for your unwavering support and for raising me right. Thanks to my Texan brother and sister, Rick and Lauren for sharing your love of this state and making me feel at home here. Thanks to all of my dear friends outside of science, for keeping me sane and grounded during this process.

Thanks to my advisor Deborah Bolnick for your support and advice throughout my PhD. Thanks to the members of my committee for taking the time to give input on my research over the past several years. Thank you to my undergraduate mentors Rika Kaestle, Michael Muehlenbein, Della Cook, and Jean Sept for fostering my interest in human prehistory and evolution and encouraging me to pursue graduate studies. Thanks to all of my colleagues at UT and elsewhere for your valuable discussions and support.

Abstract

Investigating the Human Population History of North America Using Genomics

Austin Williamson Reynolds, Ph.D.

The University of Texas at Austin, 2018

Supervisor: Deborah Bolnick

The application of genomic technologies to the study of ancient DNA over the past 10 years has revolutionized our understanding of human prehistory across the globe. A lot of work has been done resolving the past migrations of human populations on the continental scale, but less focus has been given to understanding the population history of small regions or answering archaeological questions presented at a single site. Moreover, despite the increase in genomic datasets American indigenous and admixed populations continue to be vastly underrepresented in the genomics literature, where ~95% of studies focus on either European or East Asian populations. A growing number of researchers are applying genomics techniques with modern and ancient populations in the Americas to fill in these gaps and answer exciting questions on these smaller geographic scales. This dissertation uses genome-wide data collected from indigenous populations in North America to address a number of regional and local questions. Ancient genomic data is used to address the population dynamics through time at the archaeological site of Xaltocan in central Mexico. Genomic data from modern populations is then used

to understand the effects of geography and colonial history on genetic diversity within Mexico. Finally, using genomic data from populations around North America, I explore the evidence for adaptation of these groups to their various environments.

Table of Contents

List of Tables

List of Figures

Introduction

The application of genomic technologies to the study of ancient DNA over the past 10 years has revolutionized our understanding of human prehistory across the globe. A lot of work has been done resolving the past migrations of human populations on the continental scale (Reich et al. 2012; Skoglund et al. 2015; Rasmussen et al. 2014, 2015; Raghavan et al. 2015; Scheib et al. 2018), but less focus has been given to understanding the population history of small regions or answering archaeological questions presented at a single site. Moreover, despite the increase in genomic datasets, American indigenous and admixed populations continue to be vastly underrepresented in the genomics literature, where ~95% of studies focus on either European or East Asian populations (Popejoy and Fullterton 2016). A growing number of researchers are applying genomics techniques with modern and ancient populations in the Americas to fill in these gaps and answer exciting questions on these smaller geographic scales (Valverde et al. 2016; Lindo et al. 2017).

This dissertation uses genome-wide data collected from indigenous populations in North America to begin answering a number of important questions at these more regional and local levels, such as about the genetic effects of the rise and fall of powerful political states, how the genetic effects of European colonialism vary geographically within a single country, and how indigenous groups have adapted genetically to their local environments. While these questions are of

particular interest to researchers and communities in North America, they have important implications for research into related topics around the world.

Chapter 1 analyzes the demographic effects of the rise and fall of powerful political states in pre-Hispanic Central Mexico over the past 2000 years using genomic data collected from 19 ancient residents from the archaeological site of Xaltocan located in the State of Mexico, north of modern-day Mexico City, dating to Aztec and pre-Aztec periods. Using these data, I test several hypotheses about the relationships between different groups that called Xaltocan home during the same time period and how the genetic structure may have changed through time at the site. I also compare the ancient DNA data to modern populations in Mexico in an attempt to identify their closest living relatives today. This work is an important case study for helping to answer questions posed by archaeologists and historians in many contexts around the world. For instance, did the rise and fall of powerful states lead to large movements of people in or out of conquered regions? In large multicultural centers, how did the residents interact with immigrants to the cities? Researchers have traditionally used historical and archaeological evidence to address such questions, and those bodies of data have provided many important insights and hypotheses. Ancient DNA data provides an important new type of data and, in many cases, can help test hypotheses derived from the historic and archaeological records.

Chapter 2 centers on understanding the genetic effects of Spanish colonization across Mexico, using a dataset of genome-wide data from nearly 1000 contemporary Mexican individuals from across the country, including newly collected data from 96 modern residents of three towns named Xaltocan (Jaltocan in one case) in central Mexico. With this dataset, I test hypotheses about the population structure within Mexico as a whole, how genetic diversity and recent admixture is geographically structured across the country, and the relationships between the residents of these three towns. I also attempt to build a demographic model to estimate the parameters of the genetic bottleneck that likely occurred in this area after European colonization. This study will increase our understanding of pre-Hispanic genetic diversity within Mexico and shed light on how Spanish interactions with indigenous groups varied across the country. This will help us learn more about the history of colonialism in Mexico and provide a valuable comparative dataset for understanding the genetic effects of colonial interactions across the Americas.

Chapter 3 combines genome-wide data from the modern residents of Xaltocan (State of Mexico) with genome-wide datasets from two other indigenous North American populations (the Inupiat of northern Alaska and several interrelated communities in the southeastern United States) to examine signals of natural selection across the continent. There are many well-characterized examples of selection in human populations around the globe (Tishkoff et al. 2015); however,

almost no studies of selection have been conducted in North America. The continent is home to a wide variety of environments, from the arctic tundra of Alaska and northern Canada, to the deserts of the southwestern United States, and the jungles of the Yucatan Peninsula in Mexico. People have lived in these environments some 20,000 years, providing ample time for selective pressures to make their mark on the genomes of these populations (citation?). Using genome-wide scans for natural selection, I find evidence for adaptation to the Arctic environment in the Inupiat and evidence for adaptation of immune genes in the Southeastern US and Mexico.

Together, these chapters illustrate the incredible potential that genomic data have for answering questions about population and evolutionary history on regional and local scales. They also highlight some of the limitations in the current state-of-the-art methods. Further work in these areas will allow us to better understand not just the distant past of humanity, but also the recent events of the past few thousand years.

Chapter 1: Reevaluating the Genetic Effects of Pre-Hispanic State Expansion in Central Mexico using Paleogenomic Data

Introduction

The application of genomic technologies to the study of ancient DNA over the past 10 years has revolutionized our understanding of human prehistory across the globe. Extensive research has been undertaken to resolve questions about the past migrations of human populations on the continental scale (nicely reviewed in Skoglund and Matheison 2018), but less attention has been given to using genomics to understand population history and addressing archaeological questions in smaller regions or more recent time periods.

One such question relates to the rise and fall of political states during early historic and prehistoric time periods. During periods of growth and stability, societies may experience increases in population size and immigration from other areas; while in periods of instability or collapse, societies may experience warfare, famine, emigration, and changes to community and socioeconomic structures. These varying social contexts can have a sizable impact on the demographic and genetic makeup of a society through time, and historians and archaeologists have therefore long been interested in questions about the rise and fall of various societies in the past. The field of paleogenomics offers a powerful tool for testing many hypotheses about these events when other lines of evidence are inconclusive or contradictory. For example, several paleogenomic studies in Europe were able to help resolve

debates about the nature of the Neolithic and Bronze Age transitions, showing that large population replacements accompanied the Neolithic transition to agriculture (Skoglund et al. 2012), as well as the transition to the Bronze Age (Haak et al. 2015; Allentoft et al. 2015) across much of the continent. Another study, looking at only 10 individuals buried in an Anglo-Saxon cemetery in southern England, was able to genetically distinguish between the native Britons and the incoming Anglo-Saxons (Schiffels et al. 2016).

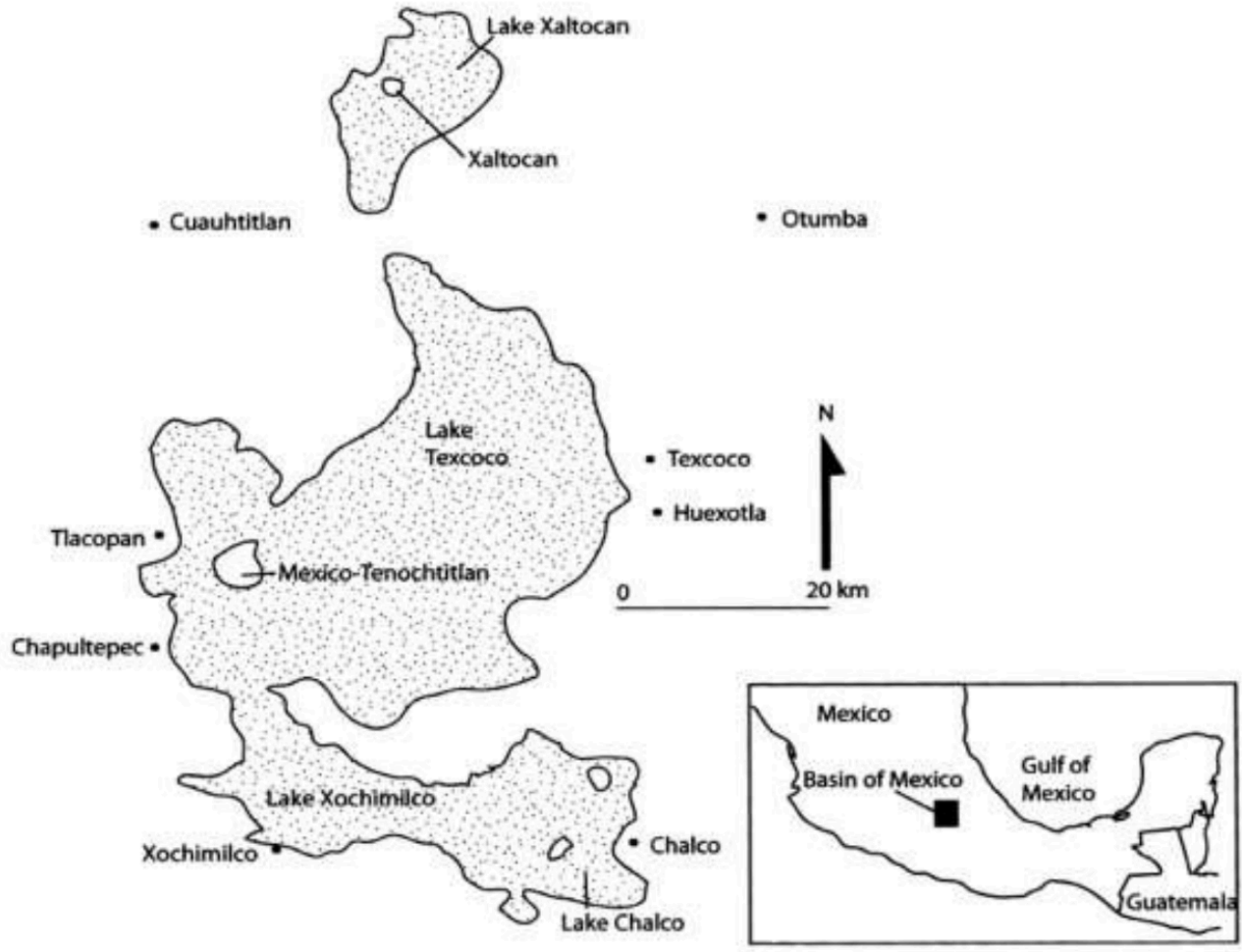
In the area around modern Mexico City, historical and archaeological research has provided a detailed picture of the successive rise and fall of the Toltec, Tepanec, and Aztec states. Following the collapse of the Toltec state in 1150 CE, various city-states in the region struggled against one another for political dominance and resources until power was eventually consolidated into a handful of large polities. By the end of the 14th century, the Tepanec state was in control of much of this region, known as the Basin of Mexico. Tepanec control was brought to an end in 1428, when an alliance between the city-states of Texcoco, Tlacopan, and Tenochtitlan formed the nascent Aztec empire. Over the next century, the Aztecs conquered nearly the entire Basin of Mexico, as well as vast areas across the broader region of central and southern Mexico. This rule lasted until the Spanish conquered the Aztec capital of Tenochtitlan in 1521 (Brumfiel 1983; Berdan and Smith 2003a,b; Smith and Berdan 2003), after which interactions between Spanish

authorities and the residents of Xaltocan seem to have been very limited (Restall and Schwaller, 2011).

While this history is well-documented, the nature and extent of the demographic and genetic effects of the political transitions in the Basin of Mexico remain unclear. For example, after the Tepanec or Aztec conquest of a neighboring polity, did the current residents continue residing in the area or were they replaced by people from the conquering group? Also, as these polities grew in size and power, did people from elsewhere immigrate to the area to take advantage of increased economic opportunity? This study presents an analysis of ~1.24 million genome-wide SNPs from 19 individuals buried at the site of Xaltocan in Central Mexico to directly test hypotheses about the genetic impacts of sociopolitical transitions in the Basin of Mexico during the Middle and Late Postclassic periods.

The town of Xaltocan, located in the Basin of Mexico (**Figure 1.1**), was founded by a group of Otomi-speaking people in 900 CE (Brumfiel 2005a). At that time, Xaltocan was a small island in the middle of the now drained Lake Xaltocan (Brumfiel 2005b; Morehart and Eisenberg 2010). From the 11th to 13th centuries, Xaltocan grew into an influential city-state that collected tribute from neighboring communities, served as the capital of the Otomi state, and experienced a high rate of population growth and land expansion, reaching a peak of around 5000 residents (Carrasco-Pizana 1987; Brumfiel, 2005a,b,c; Gibson 1964; De Lucia 2010; Sanders et al. 1979). However, by the mid-13th century, Xaltocan and the neighboring Tepanec

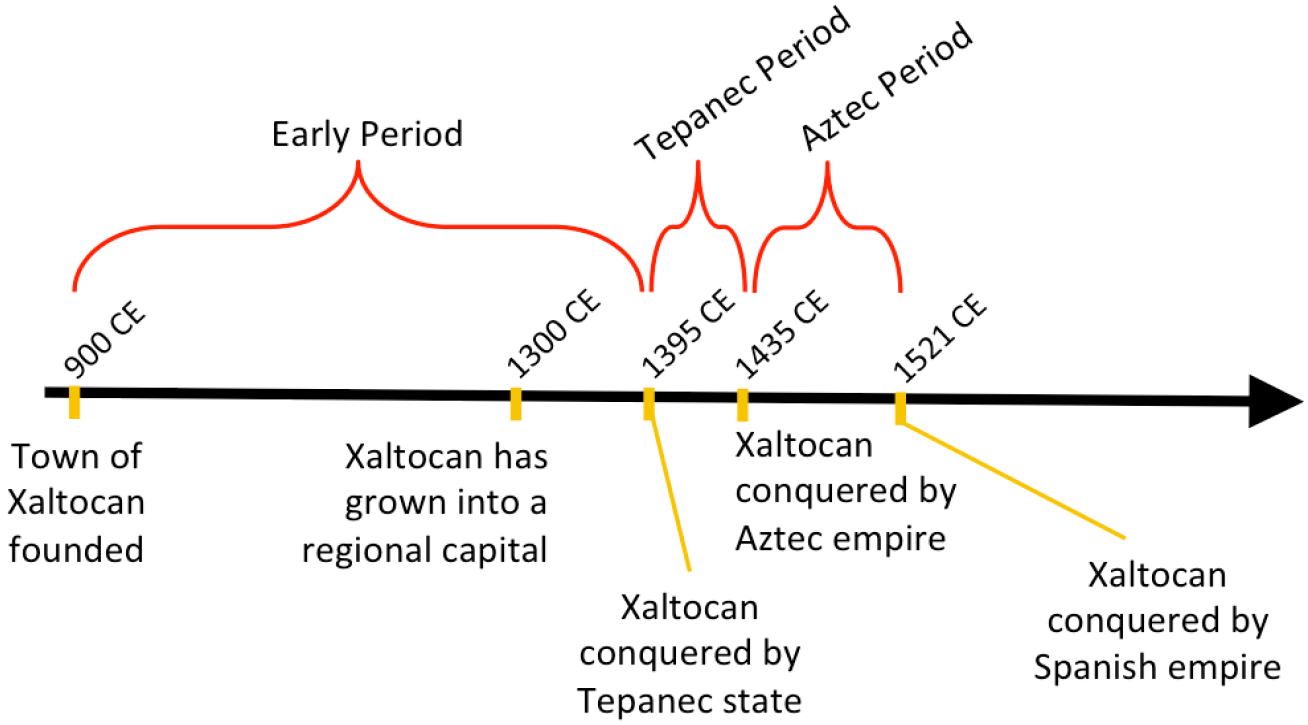
Figure 1.1 - Map of Xaltocan withdi the Basin of Mexico in ancient times (Rodríguez-Alegría, 2010)



town of Cuauhtitlan became involved in a war against each other (Bierhorst 1992; Carrasco-Pizana 1987), significantly reducing Xaltocan's influence in the Basin of Mexico. With the aid of other Tepanec allies, Cuauhtitlan eventually defeated Xaltocan in 1395. Historical accounts claim that the original Otomi-speaking inhabitants of Xaltocan abandoned the town at this time, leaving it uninhabited for 40 years (Ixtlilxochitl 1975, 1977). The emerging Aztec empire annexed Xaltocan in 1428 and, according to colonial records, assigned rulers to govern the town who repopulated it with people from other areas (Carrasco-Pizana 1987; Brumfiel 2005b; Hicks 1994). Xaltocan remained under Aztec rule until 1521, when Hernán Cortéz and his troops incorporated it into the Spanish empire (see **Figure 1.2** for a brief timeline of events at Xaltocan and period names).

While the history recorded in colonial accounts seems clear-cut, recent archaeological work suggests a somewhat different history, perhaps indicating a significant degree of continuity at Xaltocan across the Tepanec conquest and Aztec annexation. Radiocarbon analyses indicate that five burials at the site likely date between 1395 and 1435, which calls into question the assertion that Xaltocan remained uninhabited during that period (Overholtzer 2012). In addition, during the Aztec period, houses were built and burials were interred in the same locations as those used by the earlier Otomi residents (Overholtzer 2015), suggesting some level of cultural continuity between the Aztec residents of Xaltocan and the previous residents.

Figure 1.2 – Brief timeline of periods and events at Xaltocan



Differences in house structure, burial practices, and ceramic consumption have also led some archaeologists to hypothesize that two ethnically distinct groups came to inhabit Xaltocan during the period before the Tepanec conquest as the town experienced population growth and land expansion. Specifically, De Lucia and Overholtzer (2014) have suggested that while the descendants of the original Otomies lived in the interior of the island, a group of newcomers settled on the eastern and southeastern periphery. Archaeologists have also sought to understand if and how burial practices differed across neighborhoods during the Early Period or from one time period to the next, and if individuals buried in close proximity at the site were closely related.

The disagreement between the historical and archaeological records regarding this series of events at Xaltocan raises interesting questions about the demographic impact of the rise and fall of ancient polities in the Basin of Mexico. In this study, I utilize genomic data collected from the ancient residents of Xaltocan to test the following hypotheses about the history of the town:

1. Genetic differences will be evident between the residents of the interior (initial) and periphery (immigrant) neighborhoods of Xaltocan, consistent with the presence of two different ethnic groups present in the town when it was at its height of power during the Early Period.

2. Significant genetic differences will be seen between the residents of Xaltocan during different time periods, suggesting that population replacements accompanied the Tepanec conquest and Aztec annexation of Xaltocan.
3. Close genetic relationships will be identified between individuals buried within the same dwelling at Xaltocan, indicative of nuclear families or familial lineages being buried in the same place.
4. A close genetic relationship will be found between the ancient residents of Xaltocan and the residents there today, suggesting continuity through time in the town.

Testing these hypotheses will not only provide a better understanding of the demographic dynamics at Xaltocan during these periods, but will also increase our understanding of the expansion and subsequent retraction of the Tepanec and Aztec states across the region. Previous genetic work has addressed these questions using low resolution mitochondrial DNA (mtDNA) and Y-chromosomal STR datasets (Mata-Míguez et al. 2012; Mata-Míguez 2016). The purpose of this chapter is to bring genome-wide data to bear on these questions for the first time and to enhance our understanding of the history of Xaltocan.

Background

Relationship between the Early Period Interior and Periphery Burials

Previous work on mtDNA, Y-chromosomes, and autosomal STRs demonstrates that individuals from the two Early Period spatial subgroups (interior

vs. periphery Xaltocan) were not closely related maternally. All eight individuals from Early Period interior Xaltocan (100%) belong to haplogroup A, whereas Early Period periphery Xaltocan includes three individuals (21%) belonging to haplogroup A, seven (50%) to haplogroup B, and four (29%) to haplogroup D (Mata-Míguez 2016). Exact tests of population differentiation based on mtDNA haplogroup and haplotype frequencies indicate that individuals from these spatial groups are unlikely to have been drawn from the same biological population ($P = 0.001$ and $P = 0.003$, respectively). No mtDNA haplotypes are shared between individuals from different spatial groups, and median-joining haplotype networks indicate that the haplotypes present in interior Xaltocan are not closely related to the haplotypes present in periphery Xaltocan. Finally, multidimensional scaling (MDS) plots based on mtDNA haplotypes suggest a high degree of genetic differentiation between Early Period interior and periphery Xaltocan. Some mtDNA haplotypes are shared by individuals from two different houses in the same area of the town, but no haplotypes are shared across neighborhood lines. Y-chromosome data was only recovered in three individuals from the Early Period buried within a single house, making a comparison between interior and periphery groups impossible. Similarly, poor data quality in the autosomal STRs from the Early Period did not facilitate a robust comparison between the interior and periphery neighborhoods.

These results are consistent with the hypothesis that residents living in the periphery of Xaltocan were an ethnically distinct group that moved to Xaltocan

during the period of land expansion. This migration would have occurred during the period of political instability that followed the fall of the Toltec state, and may have contributed to population growth at Xaltocan and the rise of Xaltocan as an influential city-state in the Basin of Mexico (Mata-Míguez 2016).

Kin Burials Within Early Period Interior and Periphery Xaltocan

The mtDNA data show close genetic relationships between individuals within each of the two spatial subgroups (interior vs. periphery Xaltocan) during the Early Period at Xaltocan. Some mtDNA haplotypes are shared by multiple individuals from the same house, such as two infant males (ZA.2 and ZA.4, not present in the genome-wide dataset) buried in the Early Period interior house Zocalo A (See **Figure 1.3** for locations of houses). This pattern of haplotype sharing indicates the burial of maternally-related kin within the house. Similarly, a newborn female (E14.3, not present in the genome-wide dataset), 7-month old male (E14.4, not present in the genome-wide dataset), and 10-month old female (E14.5, present in the genome-wide dataset) buried close together in Structure 122 share the same mtDNA haplotype, indicating another kin burial in a house located at the periphery of Xaltocan. There are also some mtDNA haplotypes shared by individuals from two different houses in the same area of the town (e.g., related individuals in Operations G and Zocalo C in Early Period interior Xaltocan, and related individuals in Structure 122 and Operation Y in Early Period periphery Xaltocan).

Figure 1.3 - Map of the island of Xaltocan in ancient times and location of houses (reprinted from Mata-Míguez 2016).

Genetic Impacts of the Tepanec Conquest

Previous work done with mtDNA, Y-chromosomes, and autosomal STR analyses all show significant genetic differentiation between Early Period and Tepanec Period Xaltocan (Mata-Míguez et al. 2012; Mata-Míguez 2016). Early Period and Tepanec Period Xaltocan do not share any mtDNA haplotypes, and networks and MDS plots of the haplotypes from the different temporal groups show that they are not closely related. An exact test of population differentiation based on mtDNA haplotype frequencies indicates that the sampled individuals from the Early Period and Tepanec Period are unlikely to have been drawn from the same biological population ($P < 0.001$). Demographic modeling using Bayesian Serial SimCoal (Excoffier et al. 2000; Anderson et al. 2005) suggests that mutation and genetic drift alone cannot explain this difference between time periods, consistent with a hypothesis of population replacement at Xaltocan following the Tepanec conquest of the town.

Autosomal STR and Y-chromosome analyses are less informative about the relationship between Early and Tepanec period groups because the Early Period samples did not amplify well for nuclear loci (likely due to poorer DNA preservation). The results are nevertheless consistent with population discontinuity following the Tepanec conquest of Xaltocan. Kingroup analysis (Konovalov et al. 2004) rules out 1st-2nd degree relationships between any Early Period/Tepanec Period pairs. Kingroup calculates the likelihood of a particular genetic relationship

for a pair of individuals (e.g., parent-offspring, siblings, half-siblings, cousins, unrelated) given their genotypes and population allele frequencies. Similarly, analyses of Xaltocan Y-chromosomes indicate substantial differences in Y-chromosome haplogroup frequencies between the two temporal populations. Haplogroup Q-L54(xM3) is common in the Early Period (67%) but much rarer in the Tepanec Period (11%), while haplogroup Q-M3 shows the opposite pattern (33% in the Early Period versus 89% in the Tepanec Period). An exact test of population differentiation based on Y-chromosome haplogroup frequencies fails to reject the null hypothesis that Early Period and Tepanec Period individuals could have been drawn from the same population (P = 0.090), probably due to the small sample sizes (n=6 and n=9, respectively).

Overall, these uniparental results suggest that the Tepanec conquest was associated with a substantial genetic replacement at Xaltocan. However, it should be noted that the Tepanec Period sample is limited to 8 individuals interred in two households located in the periphery neighborhood of the site; those are the only individuals from the Tepanec Period that have been excavated to date. That, along with the fact that Mata-Míguez and colleagues were only using a handful of markers, suggests that more data are needed to clarify the extent of the genetic shift at Xaltocan following the Tepanec conquest.

Genetic Impacts of Aztec Annexation

Analysis of the Tepanec to Aztec transition at Xaltocan has been limited because we only have Aztec Period samples from a few individuals and a single household at the site. Previous analysis of the Aztec residents showed that they exhibited distinct mtDNA haplotypes from the Tepanec residents of the town, indicating that none of the Aztec individuals were directly descended from any of the Tepanec females studied to date (Mata-Míguez et al. 2012; Mata-Míguez 2016). Kingroup analyses of autosomal STR genotypes also rules out a 1st or 2nd degree relationship between any Aztec individual and any Tepanec individual. Analyses of Y-chromosome haplotypes also suggest that the single Aztec male sampled did not directly descend from any of the sampled Tepanec male residents of the town. Overall, the mtDNA, autosomal, and Y-chromosome evidence suggest that the Aztec period residents of this house were not closely related to the earlier Tepanec residents. This finding is consistent with the suggestion in early colonial documents that an Aztec governor populated Xaltocan in 1435 with people who were not descendants of the earlier residents of the town.

Kin Burials Within Tepanec and Aztec Houses

Previous work also found genetic relationships between individuals within both the Tepanec time period and the Aztec time period. Both Kingroup (Konovalov et al. 2004) and Cervus 3.0 (Kalinowski et al. 2007) were used to assess potential

parentage relationships between 17 individuals recovered in undisturbed, primary burials in the Structure 122 house mound (Mata-Míguez 2016). The Cervus program applies a maximum likelihood approach to evaluate how likely a female and a male are to be the parents of a given individual. Cervus results suggest the presence of a nuclear family buried together during the Tepanec Period: an adult female (E8.2, present in the genome-wide dataset) and an adult male (E8.1, present in the genome-wide dataset) were the parents of a male newborn (E8.4, not present in the genome-wide dataset) and two adult females (E8.5 and E10.2, first is present in the genome-wide dataset, the second is not). Kingroup results are consistent with a father-son relationship between E8.1 and E8.4 ($P < 0.050$) as well as with full sibling relationships between E8.5 and E10.2 ($P < 0.050$), between E8.5 and E8.4 ($P < 0.050$), and between E10.2 and E8.4 ($P < 0.050$). In addition, Y-chromosome haplotype data confirm that E8.1 and E8.4 were closely related paternally, consistent with a father-son relationship between these individuals. E8.2, E8.4, E8.5, and E10.2 share the same mtDNA haplotype, indicating that they are all maternally related. Because a 5-year old male (E7.1, present in the genome-wide dataset) and an adult female (E8.3, not present in the genome-wide dataset) buried nearby also exhibit the same mtDNA haplotype, this points to the presence of an extended family and provides evidence that six individuals who belonged to the same matriline lived, died, and were buried in a single structure during the 14th and 15th centuries, probably over multiple generations.

Structure 122 also exhibits a cluster of three individuals interred on the eastern side of the patio during the Aztec period who did not appear closely related to the residents of the western side during the Tepanec period. Kingroup indicates that two of the individuals on the eastern side of the patio, an adult female (E14.6 present in the genome-wide dataset) and a 14-year old male (E6.1, present in the genome-wide dataset), were likely mother and son. In accordance with this inference, E14.6 and E6.1 share the same mtDNA haplotype.

Relationship of Ancient Xaltocan to Modern Populations in Mexico

Previous work compared Late Period (Tepanec, Aztec) mtDNA and Y-chromosome data to that from residents of modern Xaltocan (Mata-Míguez et al. 2016). In the Late Period, ten individuals (63%) belong to haplogroup A, three individuals (19%) belong to haplogroup B, one individual (6%) belong to haplogroup C, and two individuals (13%) belong to haplogroup D. These mtDNA haplogroup frequencies are consistent with those found in previous studies of ancient and modern native populations in Mexico (Malhi et al. 2003; Sandoval et al. 2009; Kemp et al. 2010). In modern Xaltocan, 26 individuals (55%) belong to haplogroup A, 12 individuals (26%) belong to haplogroup B, and 9 individuals (19%) belong to haplogroup C. Modern Xaltocan is more diverse than pre-Hispanic Xaltocan in terms of mtDNA haplogroup and haplotype diversity. However, ancient Xaltocan is more diverse than modern Xaltocan in terms of mtDNA nucleotide

diversity. Haplotype networks show that ancient and modern Xaltocan do not share any mitochondrial haplotypes (Mata-Míguez 2016).

However, all of the modern residents belong to founding Native American mtDNA haplogroups, and all of the modern male residents with a paternal grandfather from Xaltocan belong to the founding Native American Y-DNA haplogroup Q (Mata-Míguez 2016). These results are consistent with only extremely low levels of female- and male-mediated gene flow from Europe and Africa into Xaltocan during the colonial period. This finding is in agreement with early colonial documents, which suggest that newly arrived Europeans and Africans rarely visited Xaltocan (Montúfar 1897; Hicks 2005).

The MDS plot based on F_{ST} values shows that there is a substantial degree of genetic differentiation between ancient and modern Xaltocan, and the test of population differentiation indicates that the ancient and modern Xaltocan mtDNA haplotypes are unlikely to represent the same biological population ($P < 0.001$). Coalescent-based simulations designed to identify the relevant processes leading to this difference indicated that demographic scenarios characterized by plausible parameter values (e.g., for population size and population growth) for ancient Xaltocan, realistic mutation rates, and the absence of migration are unlikely to explain the lack of shared mtDNA haplotypes between pre-Hispanic and modern Xaltocan. Because genetic drift and mutation alone do not explain the observed

genetic differences between these temporal populations, this result suggests that gene flow likely led to these genetic differences.

Oral History Suggests Possible Link Between Ancient and Modern Xaltocans

Historical accounts claim that the original Otomi-speaking inhabitants of Xaltocan abandoned the town after the Tepanec conquest, leaving it uninhabited until the Aztec empire annexed and repopulated the area 40 years later Ixtlilxochitl 1975, 1977). This had led archaeologists to wonder where the original inhabitants of Xaltocan may have gone after abandoning the town, if indeed they truly left. There are two other towns in present-day Central Mexico that share a name with Xaltocan. Oral histories collected by Dr. Jaime Mata-Míguez in the summer of 2015 in these towns (Xaltocan, Tlaxcala and Jaltocan, Hidalgo) suggest that there might be some connection between the ancient residents of Xaltocan and the ancestors of the people living in these other towns. He subsequently collected saliva samples from residents of these two towns to test if these individuals share a closer genetic relationship with the ancient residents of Xaltocan than those living there today, possibly leading to the shared names between these towns.

Results

To answer the many remaining questions about the genetic history of Xaltocan, I analyzed newly collected genome-wide data from ancient residents of the town. Ancient DNA was extracted from 38 of the skeletal samples following published protocols (Rohland and Hofreiter 2007; Bolnick et al. 2012; Dabney et al.

2013). Nineteen of the 38 extracted individuals, spanning the Early and Late Periods (N=3 Early Interior; N=6 Early Periphery; N=5 Tepanec; N=2 Aztec; N=3 Late Unknown), were selected for genome-wide analysis based on evidence for good DNA preservation. Genomic libraries were constructed (Rohland et al. 2015; Haak et al. 2015), treated with an in-solution enrichment for a targeted set of ~1.24 million SNPs (Mathieson et al. 2015) and sequenced. I also collected genome-wide genotype data on ~620K SNPs from 47 current residents of the town of Xaltocan, as well as 46 samples from current residents of two other towns, Xaltocan in the Mexican state of Tlaxcala (N=22) and Jaltocan in the Mexican state of Hidalgo (N=24) using the Affymetrix Human Origins Array. The newly collected genomic data from both the ancient and modern samples were combined with a comparative dataset of global populations genotyped on various arrays, including Native American individuals who were masked for non-indigenous ancestry. The final comparative dataset consisted of 1,045 individuals and 101,679 SNPs.

This cross-section of ancient data points, along with the large comparative dataset, allowed me to test hypotheses about the differences between the two neighborhoods during the Early Period, as well as the relationships between residents of Xaltocan across different time periods. I was able to test my hypothesis about the genetic relationships between individuals buried in the same dwelling at the site, and the new data also allowed me to attempt to assign individuals buried at the site of unknown cultural association to either the Tepanec or Aztec Period based

on genetic similarity to each group. Finally, using the data collected from the modern residents of the Xaltocans, I was able to assess the relationship of the ancient residents of Xaltocan and the people living there today, as well as test if the ancient residents of Xaltocan share a strong genetic affinity with the people living in the Xaltocans in Tlaxcala and Hidalgo, as the oral history suggests may be the case.

I first used the newly collected data to corroborate the relationships identified previously and to determine if any relationships existed that had not already been detected. I identified pairs of related ancient individuals from the genome-wide data based on the proportion of sites covered in pairs of ancient samples from the same population that had identical allele calls. I confirmed the close relationship seen previously between E6.1 and E14.6, as well as that between E8.5 and individuals E7.1, E8.2, and E8.1 (**Table 1.1)**. Individuals E8.5 and E6.1 were therefore removed from population scale analyses to avoid biasing the results.

To better understand the relationships among the ancient groups at Xaltocan as well as their relationship with present-day indigenous Mexican populations, I constructed principal components using the data from indigenous Mexican groups masked for non-indigenous ancestry, and then projected the ancient individuals onto the PCs. **Figure 1.4**, showing all ancient individuals, demonstrates that they all cluster with Central Mexican populations, including the modern residents of Xaltocan at this scale. Zooming in, **Figure 1.5** shows that the ancient Xaltocan

Table 1.1 – Relatives detected using genome-wide data. Samples in the last column are first-degree relatives of samples in the first column.

SampleID	Time Period	1st degree rels
E8.2	Tepanec	E8.5
E8.1		E8.5
E7.1		E8.5
E8.5		E8.1, E7.1, E8.2
E6.1	Aztec	E14.6
E14.6		E6.1

Figure 1.4: PCA of masked modern Mexican populations with ancient Xaltocan individuals projected.

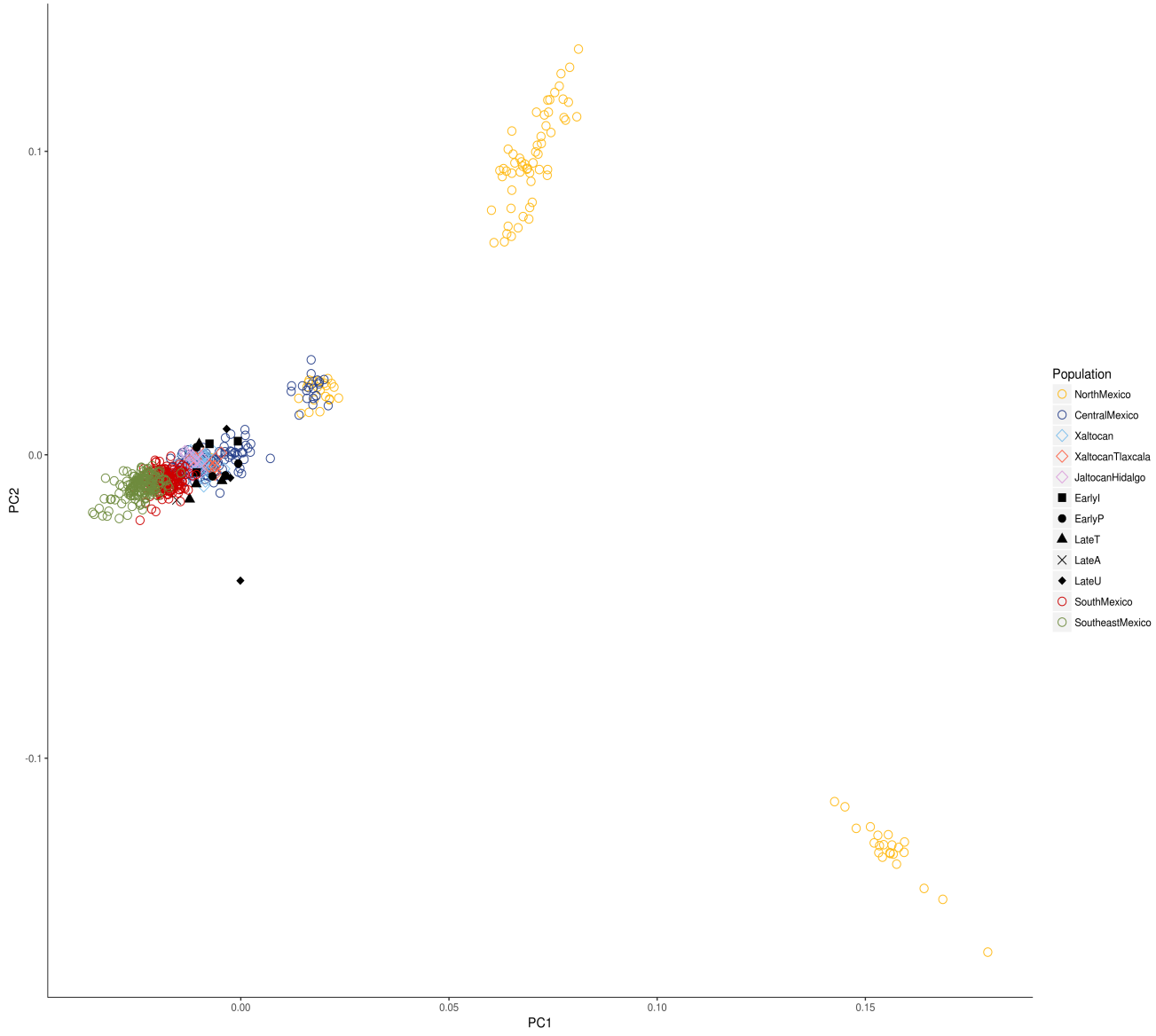
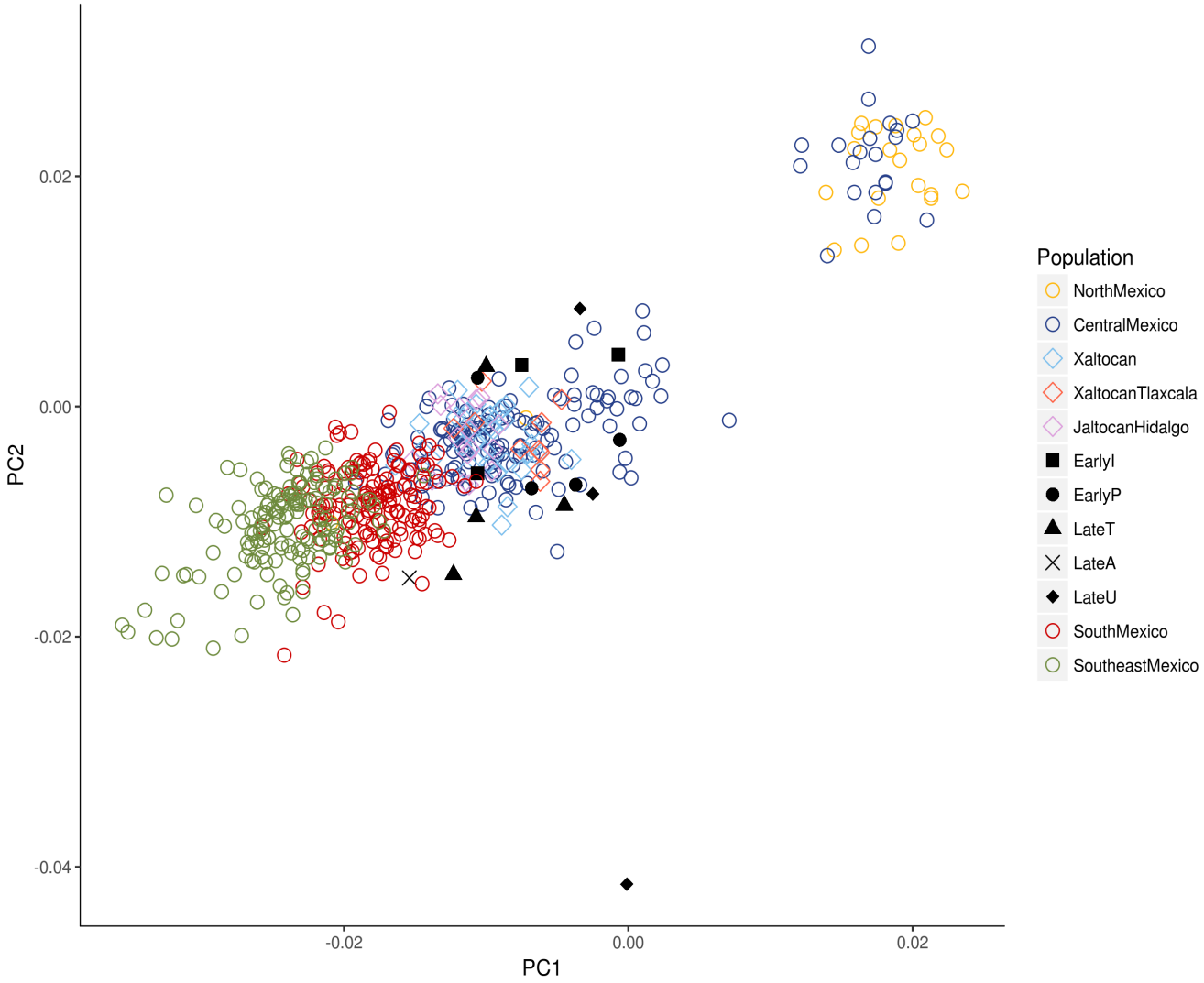


Figure 1.5: ***PCA of masked modern Mexican populations with ancient Xaltocan individuals projected, zoomed in.***



individuals tend to cluster with one another to the exclusion of the other groups. No clear distinctions are evident between the Early Period and the Tepanec period individuals at this scale, nor between the Tepanec period and the Aztec period individuals. It is also not possible to clearly assign any of the Late period individuals of unknown provenance to the Aztec or Tepanec periods using PCA analysis because the Aztec and Tepanec samples are not clearly divided into separate clusters in the PCA. I then used discriminant function analysis based on the principal component loadings to attempt to find which principal components would best allow me to tell the time periods apart. A MANOVA of the principal components identified in a discriminant function analysis revealed no significant differences between the ancient groups based on the principal components.

ADMIXTURE analysis (**Figures 1.6-7,14,15**) was next conducted to vizualize the similarities between the ancient groups and modern comparative populations. ADMIXTURE runs at K=15 showed the lowest cross-validation error and are therefore presented in **Figures 1.6 and 1.7**. This analysis shows that the ancient individuals at Xaltocan look very much like present-day individuals from Central and Southern Mexico. While there is some variability from individual to individual, there are no major differences in the ancestry proportions of any subgroup within the ancient Xaltocan samples to differentiate them from the others.

To test for significant relationships among the ancient groups at Xaltocan and between each ancient group and all of the modern groups in the comparative

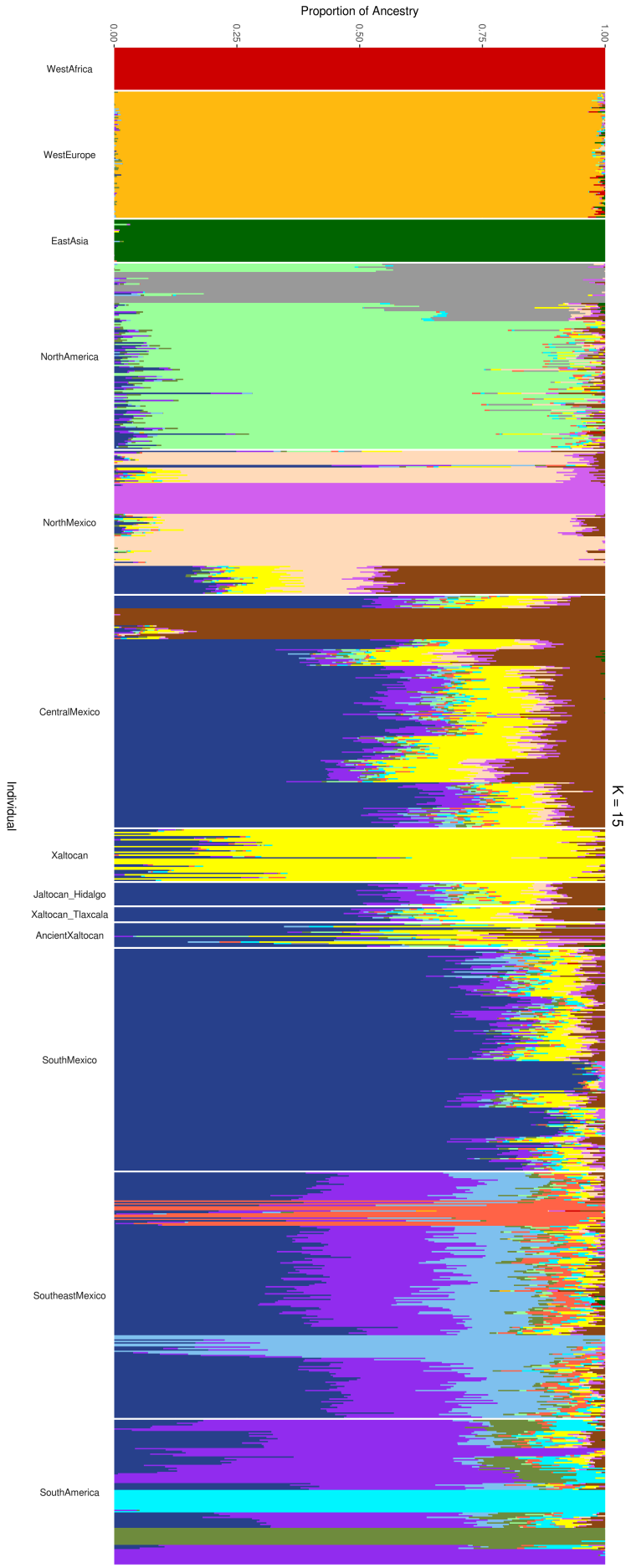
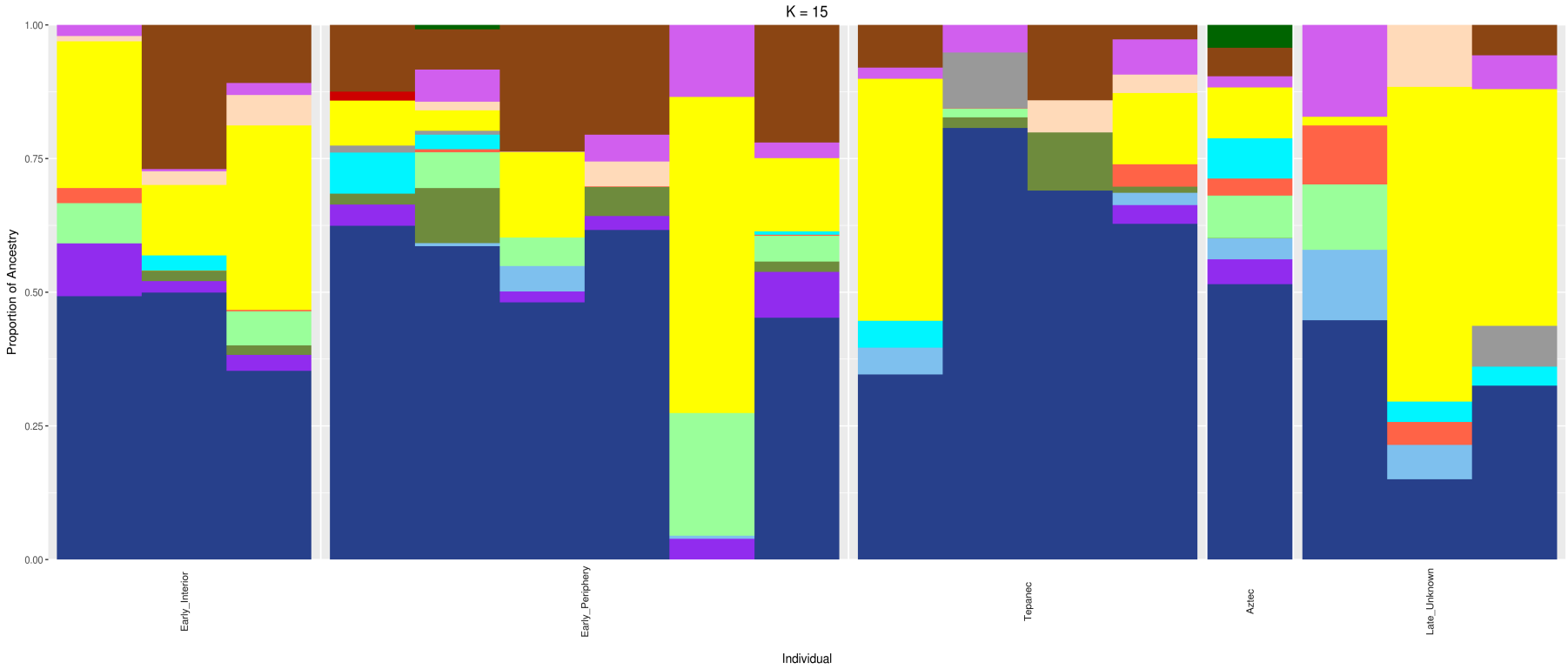


Figure 1.6 – ADMIXTURE results at K=15

Figure 1.7 – ADMIXTURE plot at K=15, Ancient Populations Only



dataset, I calculated *f*4 statistics of the form (YRI,X;Y,Z) where YRI is Yoruba, X is one of the ancient Xaltocan groups, and Y and Z are every pairwise combination of the comparative populations plus the ancient groups (Patterson et al. 2012). This test evaluates if the data are consistent with a four-population tree of the form: (Outgroup, Population X; PopulationY, PopulationZ). The value of the statistic under the null hypothesis is zero, while positive values suggest that the PopulationX is closer to PopulationZ, and negative values suggest that the PopulationX is closer to PopulationY. Values with an absolute z-score higher than 3 were considered significant. Unsurprisingly, the ancient groups at Xaltocan are significantly closer to Mesoamerican and South American groups than to populations in North America (**Table A1**), consistent with current models of the peopling of the Americas (Reich et al. 2012; Skoglund et al. 2015; Rasmussen et al. 2014, 2015; Raghavan et al. 2015; Scheib et al. 2018). Ancient groups from Xaltocan are also significantly more related to groups within Mexico than to South American populations. Within Mexico, the ancient Xaltocan groups are most genetically similar to modern populations living in Central and Southern Mexico, and most genetically distant from populations living in Northern Mexico.

Table 1.2 shows the significant *f*4 results within the Central and Southern Mexican comparative populations. The two Early Period groups appear significantly closer to the Nahuatl-Morelos population (Nahuatl speakers living in the state of Morelos) than several, but not all, of the other Central and Southern Mexican

Table 1.2 – Significant f4 results within central/southern Mexican populations. (EarlyP=Early Periphery, EarlyI = Early Interior, LateT = Tepanec)

YRI	X	Y	Z	D	Zscore	SNPs
Yoruba	EarlyP	Huasteco	NahuatlMorelos	0.0081	3.417	84429
Yoruba	EarlyP	Huichol	MixtecoAlto	0.0075	3.003	84429
Yoruba	EarlyP	Huichol	NahuatlMorelos	0.0114	4.521	84429
Yoruba	EarlyP	Huichol	TotonacoVer	0.0087	3.051	84429
Yoruba	EarlyP	NahuaPuebla	NahuatlMorelos	0.0081	3.587	84429
Yoruba	EarlyP	NahuatlMorelos	NahuaTrios	-0.0081	-3.453	84429
Yoruba	EarlyI	Huichol	NahuatlMorelos	0.0131	4.451	78868
Yoruba	EarlyI	Mazatec	NahuatlMorelos	0.0089	3.35	78868
Yoruba	EarlyI	Mixe2	NahuatlMorelos	0.011	3.424	78868
Yoruba	EarlyI	NahuatlMorelos	NahuatlSLP	-0.0117	-3.674	78868
Yoruba	EarlyI	NahuatlMorelos	NahuaTrios	-0.0095	-3.694	78868
Yoruba	EarlyI	NahuatlMorelos	Purepecha	-0.0088	-3.381	78868
Yoruba	EarlyI	NahuatlMorelos	Totonac	-0.0091	-3.374	78868
Yoruba	EarlyI	NahuatlMorelos	XaltocanTlaxcala	-0.0091	-3.071	78868
Yoruba	EarlyI	NahuatlMorelos	ZapotecSouth	-0.0082	-3.187	78868
Yoruba	LateT	Huichol	NahuatlMorelos	0.0117	3.051	39533
Yoruba	LateT	Huichol	Triqui	0.0109	3.139	39533
Yoruba	LateT	Huichol	ZapotecNorth	0.0107	3.143	39533
Yoruba	LateT	Huichol	Zapoteco	0.0146	3.284	39533

populations. Both the Early Periphery and Early Interior groups appear to be somewhat distant to the Huichol population (a group indigenous to the Sierra Madre mountains north of the city of Guadalajara). The only significant result for the Tepanec in comparison with Central and Southern Mexican populations is that they are genetically distant from the Huichol. The results with the Huichol are perhaps not surprising as they appear to be an outlier in both the ADMIXTURE and PCA plots relative to other Central and Southern Mexican groups, and they are the northern-most group categorized as Central Mexico in this analysis.

Beyond that, it is difficult to interpret these results because little genetic differentiation was detected among the groups in this region. The Aztec individuals, for instance, show no statistically significant similarity or dissimilarity with any of the nearby modern populations. While the Early Period residents of Xaltocan appear closest to the present-day Nahuatl-Morelos population, the Tepanec and Aztec residents of Xaltocan are not significantly dissimilar to that modern group compared to other populations in the dataset. Put simply, there are not clear patterns of genetic differentiation within the comparative dataset that allow me to differentiate between the ancient groups at Xaltocan with the *f4* statistics. There are no significant relationships between the ancient groups and the modern residents of Xaltocan or with the other two towns named Xaltocan that were sampled for this project. Furthermore, direct comparisons between the ancient groups do not show

significant affinities or differences among them, indicating little differentation among them at the loci studied (**Table 1.3**).

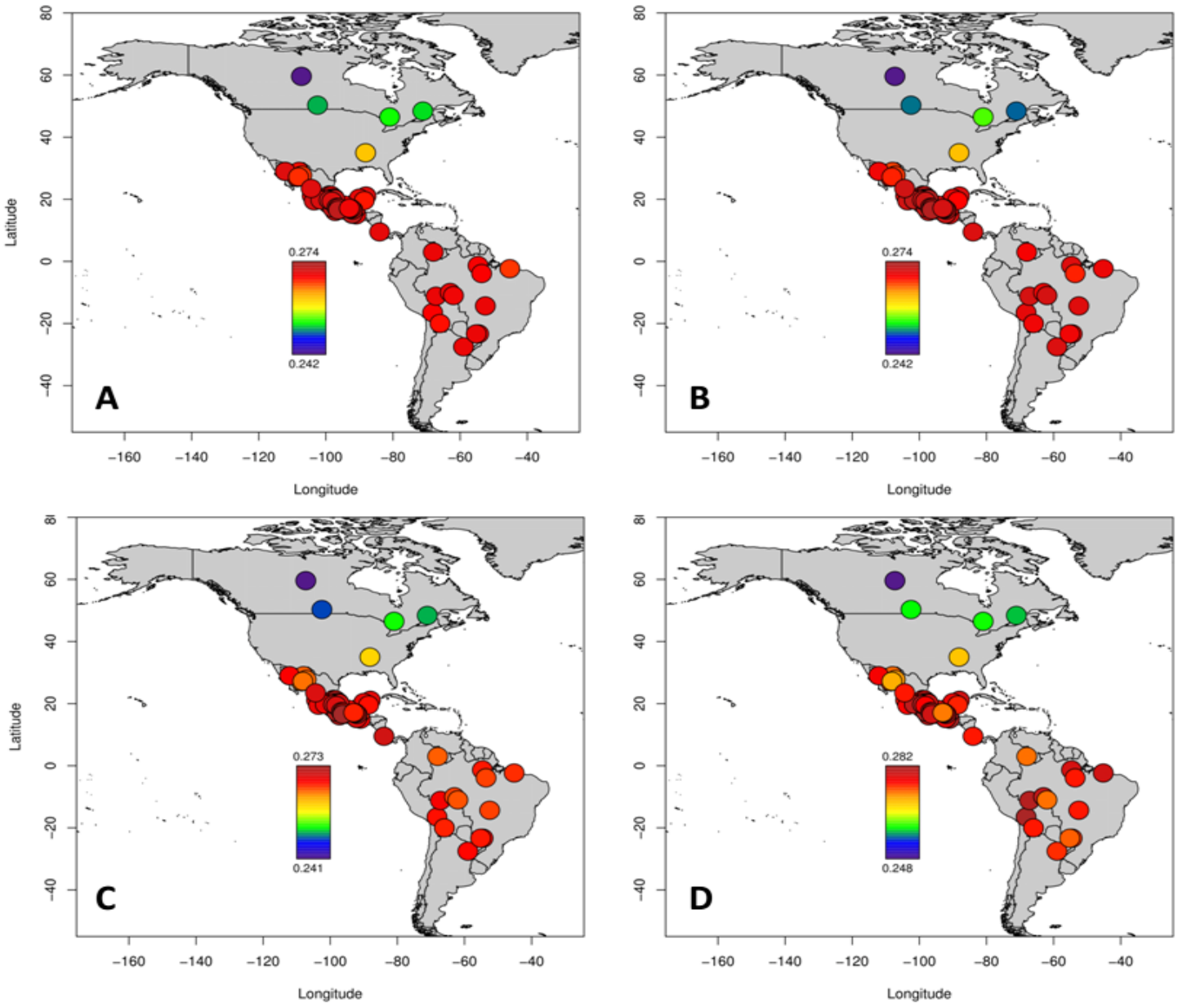
I next used *qpWave* to try to leverage the modern genetic variation within the Americas to distinguish between the ancient groups at Xaltocan. *QpWave* uses a large matrix of f4-statistics measuring allele sharing correlation rates between all possible pairs of a set of outgroups (a number of modern American populations) and all possible pairs of a set of test populations (the various ancient Xaltocan groups). *QpWave* then determines whether allele frequencies in the test populations can be explained by one or more streams of ancestry derived in different ways from the outgroups. Here, *QpWave* was not able to significantly find more than one stream of ancestry into all of the ancient Xaltocan groups, meaning that it was not able to distinguish between the groups.

I also used outgroup-*f*3 statistics (Raghavan et al. 2014) of the form (X,Y;Yoruba, where X is an ancient Xaltocan group and Y is a population from the comparative dataset) to get a better idea of where in Mexico and around the Americas the closest relatives to each ancient group at Xaltocan live today. The value of the outgroup-*f*3 statistic is proportional to the relatedness between Population X and Population Y, and therefore, greater values mean that Population X and Population Y share a higher degree of genetic drift. **Figure 1.8** shows that the ancient residents of Xaltocan exhibit greater genetic affinity to modern Mesoamerican and South American populations than to North American groups.

Table 1.3 – f4 results for Late Period individuals of unknown culture.

YRI	X	Y	Z	*f*4	Zscore	SNPs
Yoruba	E14.1	EarlyI	EarlyP	-0.0095	-0.264	2038
Yoruba	E14.1	EarlyI	LateA	0.0631	0.686	545
Yoruba	E14.1	EarlyI	LateT	-0.0604	-1.215	1319
Yoruba	E14.1	EarlyP	LateA	0.0265	0.317	548
Yoruba	E14.1	EarlyP	LateT	-0.0494	-0.955	1336
Yoruba	E14.1	LateT	LateA	0.0509	0.433	430
Yoruba	E30.3	EarlyI	EarlyP	0.0111	0.76	11933
Yoruba	E30.3	EarlyI	LateA	-0.02	-0.588	3632
Yoruba	E30.3	EarlyI	LateT	0.0105	0.489	8064
Yoruba	E30.3	EarlyP	LateA	-0.0271	-0.83	3655
Yoruba	E30.3	EarlyP	LateT	-0.0051	-0.251	8158
Yoruba	E30.3	LateT	LateA	0.0088	0.199	2782
Yoruba	E34.1	EarlyI	EarlyP	-0.0149	-1.266	19534
Yoruba	E34.1	EarlyI	LateA	-0.0366	-1.213	4662
Yoruba	E34.1	EarlyI	LateT	-0.0227	-1.303	11943
Yoruba	E34.1	EarlyP	LateA	0.0058	0.212	4700
Yoruba	E34.1	EarlyP	LateT	-0.0002	-0.013	12155
Yoruba	E34.1	LateT	LateA	0.0032	0.081	3439

Figure 1.8 – Outgroup-f3 results for each subgroup of the ancient Xaltocan samples (A. Early Interior, B. Early Periphery, C. Tepanec, D. Aztec)



This is in line with previous results reported elsewhere (Reich et al. 2012; Raghavan et al. 2015). Within Mexico (**Figures 1.9-1.12**), the ancient Xaltocan groups are closest to people living in Central and Southern Mexico today, in line with the results seen in the *f*4 statistics. Each of the different ancient Xaltocan groups is "closest" to a different modern population, but taking into account confidence intervals for the statistics, it appears that the ancient and modern residents of this region are all very closely related. Because of variable sequencing coverage in the ancient data, I also ran pairwise outgroup-f3 statistics between each pair of ancient individuals (in the form of X,Y;Yoruba, where X and Y are pairs of ancient individuals; **Figure 1.13**). If there is more genetic differentiation between groups than between individuals within groups, the outgroup-*f*3 statistics should be higher between within-group pairs than between inter-group pairs. However, there is no clear pattern between groups, suggesting that any significant genetic differentiation between the ancient subgroups at Xaltocan with the outgroup-*f3* statistics.

Discussion

Understanding the demographic and genetic effects of the rise and fall of past polities remains an important area of inquiry in bioarchaeological research. Paleogenomics has proven very useful for helping to test various hypotheses about social organization and migration patterns in the past (Skoglund et al. 2012; Haak et al. 2015; Schiffels et al. 2016; Amorim et al. 2018). This study provides important new evidence regarding population dynamics during the Middle and Late Preclassic

Figure 1.9 – Outgroup-f3 results within Mexico for the Early Interior. Samples show close relationships with modern Central and Southern Mexico populations

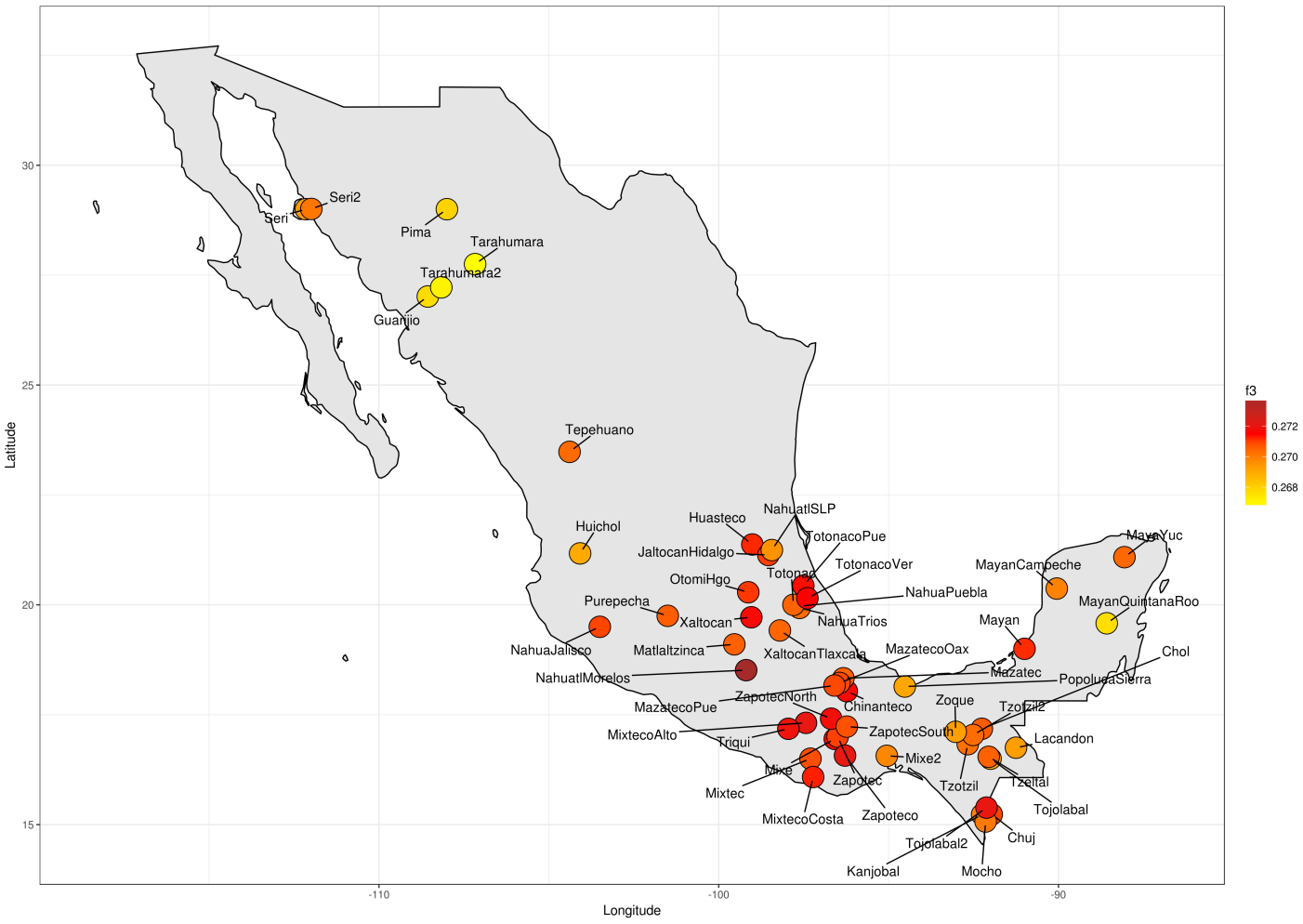


Figure 1.10 – Outgroup-f3 results within Mexico for the Early Periphery. Samples show close relationships with modern Central and Southern Mexico populations

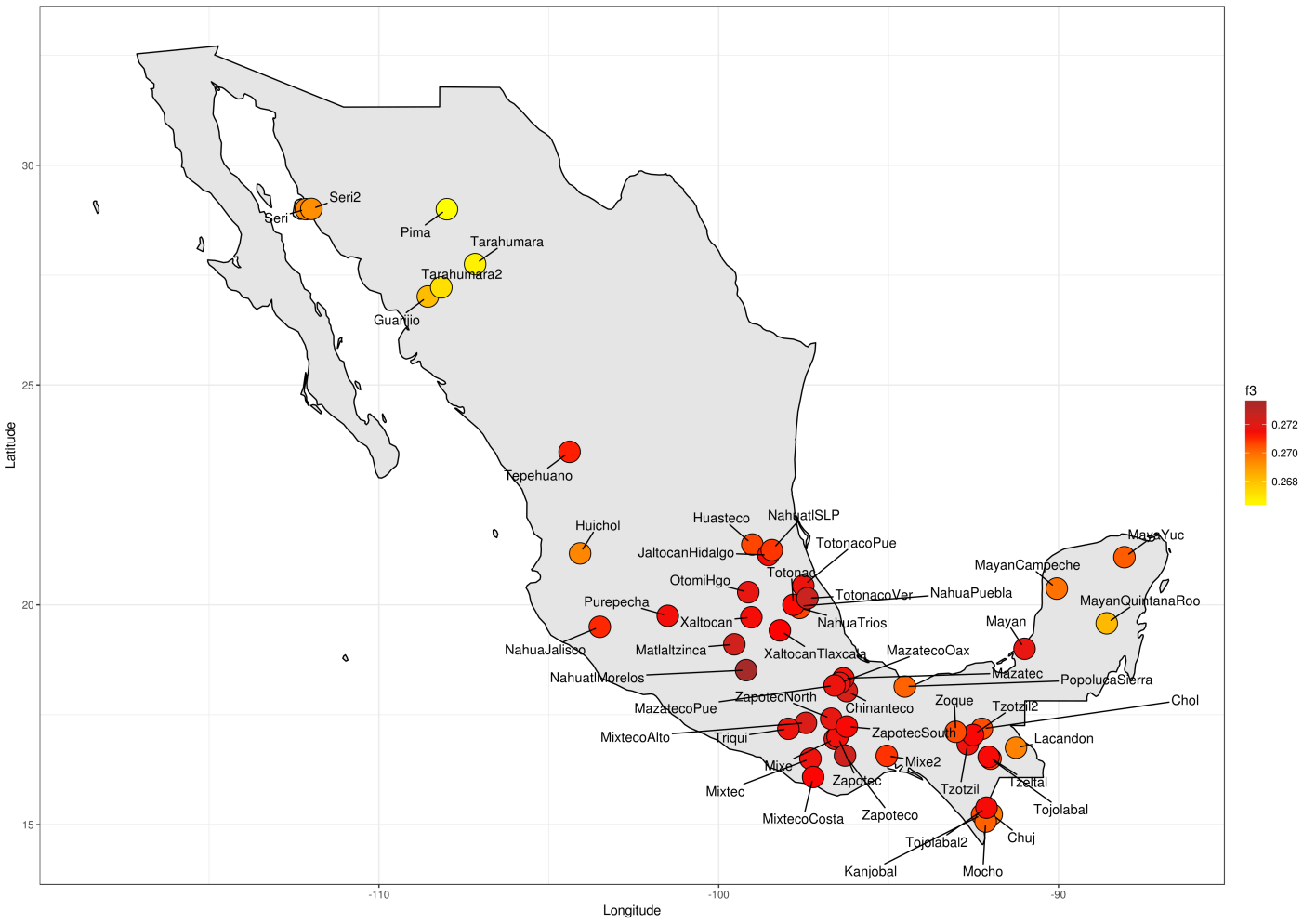


Figure 1.11 – Outgroup f3 results within Mexico for the Tepanec. Samples show close relationships with modern Central and Southern Mexico populations

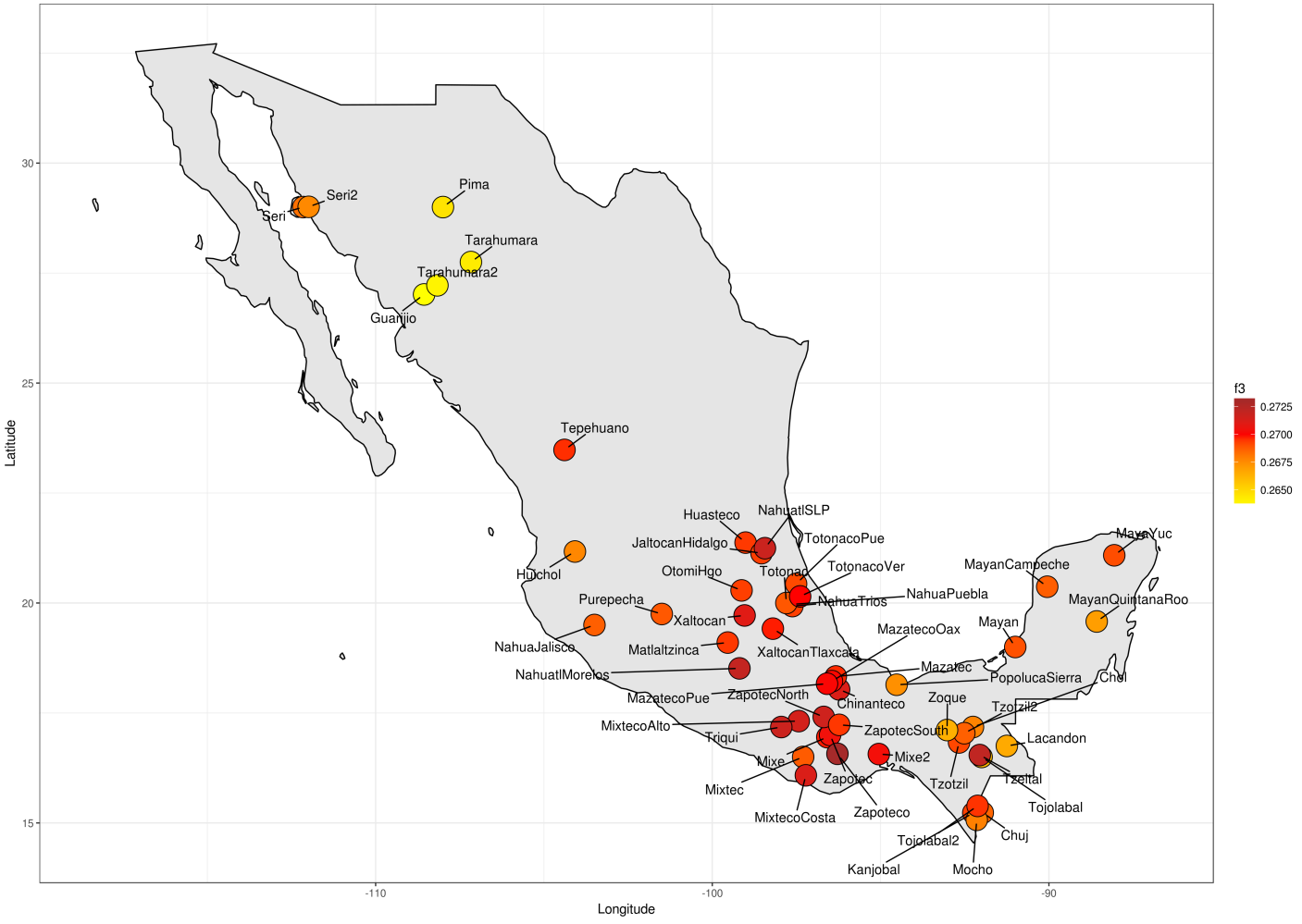


Figure 1.12 – Outgroup f3 results within Mexico for the Aztec Samples show close relationships with modern Central and Southern Mexico populations

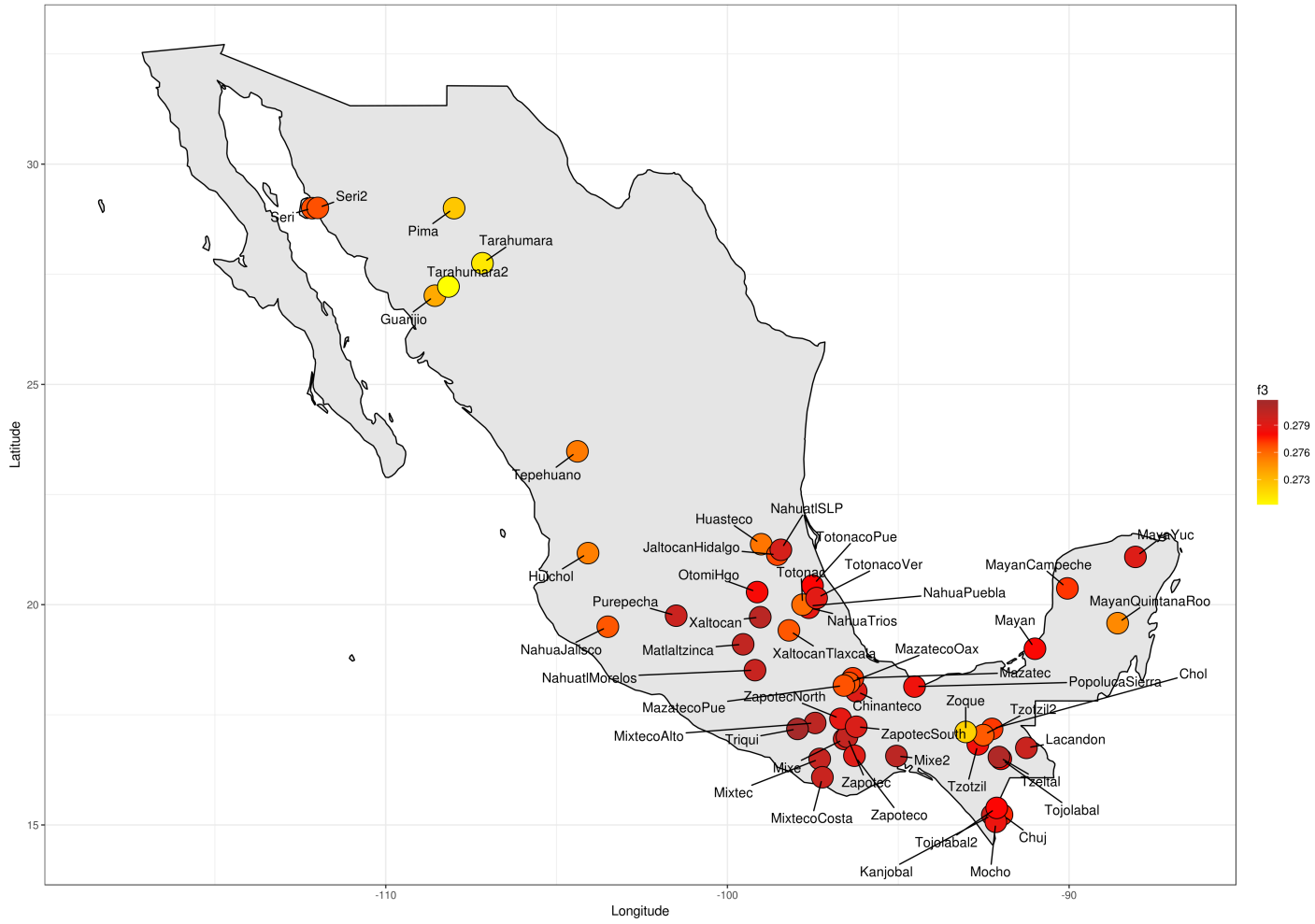
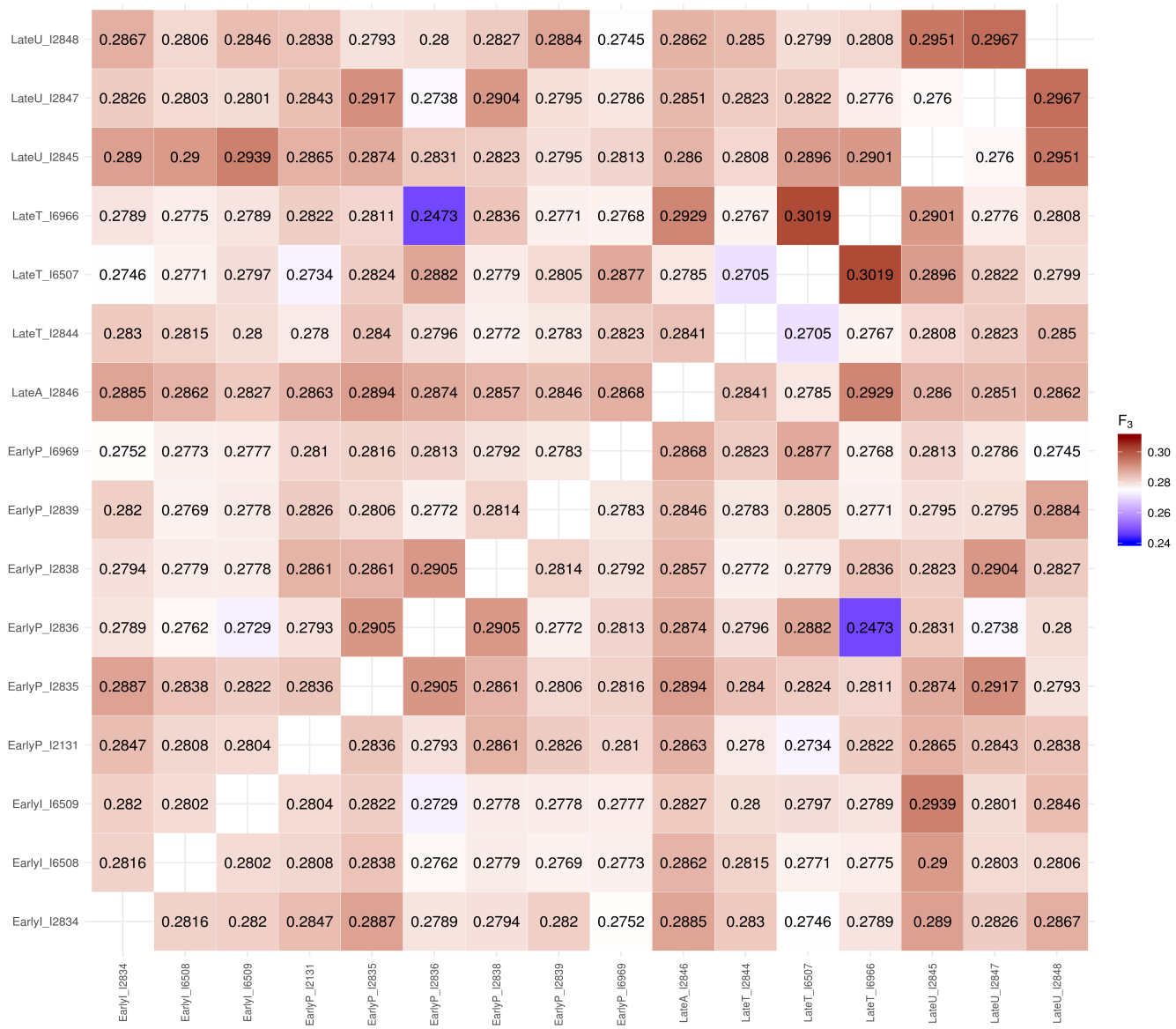


Figure 1.13 – Pairwise outgroup-f3 statistics



period in Mexico, immediately prior to and during the formation of the Aztec empire.

The results of genome-wide analysis of individuals interred during the different periods at Xaltocan provide mixed support for the proposed hypotheses. The genome-wide data collected here provide no evidence of genetic differentiation between the Early Period Interior and Periphery neighborhood residents. Nor do these new data offer support for genetic differentiation between the Early and Tepanec Periods or between the Tepanec and Aztec Periods at the site. This is in contrast to the stark differences in mtDNA and Y-chromosome haplotypes seen between neighborhoods and time periods at the site.

The newly collected data do allow us to gain a better understanding of the relationships between the ancient residents of Xaltocan and populations living in Mexico today. All of the ancient individuals at Xaltocan show a close genetic relationship to populations in Central and Southern Mexico. However, they are not significantly more closely related to the contemporary residents of the three Xaltocans than other populations in Central and Southern Mexico. This is in line with the historic record as well as the results from the analysis of uniparental data (Y-chromosome and mtDNA) of Mata-Míguez (2016), which suggest that a lot of gene flow occurred between groups in this region both before and during the colonial period. The results of this analysis do not support the hypothesis that the current residents of Xaltocan, Tlaxcala and Jaltocan, Hidalgo are the descendants of people

who abandoned Xaltocan after the Tepanec conquest. The oral histories of the residents of those towns may therefore point instead to migrants coming to their communities from another source or sources. Finally, it should be noted that the newly collected genome-wide data do confirm all of the first-degree relationships between ancient individuals at the site that were suggested from the uniparental and autosomal-STR data.

Study Challenges

There are several possible explanations for the discordance between the results of the genome-wide analyses and the earlier findings based on uniparental markers. I suggest three possible reasons here, though it is likely a combination of all of them. (1) The small sample size within each time period for the genome-wide analysis, particularly the single individual available from the Aztec period, may be limiting the statistical power to detect differences between groups at the site. Small sample size has not been an issue in similar paleogenomic studies in Europe (Schiffels et al. 2016; Skoglund et al. 2014), but small sample size coupled with points (2) and (3) below has likely contributed to the results seen in this study. (2) I used the 1240K capture, which ascertained for commonly variable SNPs worldwide, to recover genetic data from the ancient residents of Xaltocan, instead of whole-genome capture or shotgun sequencing. Therefore, I may not be obtaining data from the loci that are most variable within the closely related indigenous populations of Mexico that would allow me to differentiate between them. This is because many

arrays are biased towards SNPs that are variable in European populations (Lachance and Tishkoff 2013). Even on the Human Origins Array, where roughly half of the SNPs in the 1240K capture are sourced (Mathieson et al. 2015), only ~12K SNPs on the array were found to be variable in a single South American population. While the other SNPs may be variable in American populations as well, such a small number of variable SNPs ascertained in the Americas biases the 1240K capture against detecting variation in these groups.

(3) As evidenced by the *f3* and *f*4 results, modern populations within Central and Southern Mexico are very closely related to one another. The ancient Xaltocan individuals are also very closely related to these populations, with no demonstrative pattern of being closest to a single population in the region, limiting the ability of current state-of-the-art analyses to detect differences between the residents of Xaltocan at different time periods. The *f*-statistics were originally designed to detect introgression between groups that have been diverged for quite some time, such as that between Neanderthals and modern humans (Green et al. 2010). They have since been used successfully to detect human population movements across the world (Skoglund et al. 2012; Moorjani et al. 2013). Methods such as qpWave have also been developed to help understand admixture in contexts where two groups had previously diverged for at least a few thousand years (Haak et al. 2015). These tools have become the state of the art in detecting population replacement and admixture in human prehistory. In the case of Xaltocan, however, it is unlikely that

any populations who sequentially occupied the site had previously diverged genetically for such a long period of time. Because of the close relationships between the populations of Central and Southern Mexico, these state-of-the-art tools may be unable to detect whatever genetic differentiation is present between the ancient groups at Xaltocan.

Given the extent of discontinuity between mtDNA haplotypes in the different periods at Xaltocan, I would expect to see some pattern of differentiation in the genome-wide data as well. However, without access to higher-coverage data from the ancient individuals and potentially new analytical methods that are more sensitive to much more recent genetic divergences, it may not be possible to confirm the patterns seen in the mtDNA and Y-chromosome data. Importantly, this does not invalidate the conclusions drawn from these uniparental data. Because of their properties (mostly non-recombining, relatively high mutation rates, etc…), mtDNA and Y-chromosomal data are uniquely situated to answer questions about recent population interactions between closely related populations. Indeed, further work to recover complete mitochondrial genomes and additional Y-chromosome data may be more beneficial for answering the questions we have in cases such as Xaltocan because they will show us genetic changes on a shorter time scale.

Possible Paths for Future Work

Very recent demographic history can be assessed with high-coverage and high-quality genomic datasets. In particular, leveraging variation in phased

haplotypes (Hellenthal et al. 2014) and rare variants within whole genome sequencing datasets has proved useful in disentangling very recent demographic changes and selective pressures (Field et al. 2016). The type of data necessary for these types of analyses is fairly straightforward to produce in modern datasets (if a bit costly); however, the majority of ancient samples are not well preserved enough to produce this kind of data. Nevertheless, recent work has shown that it is possible to use rare variants in ancient datasets to look at very fine-scale demographic structure between closely related groups (Schiffels et al. 2016). Further work along these lines, and possibly imputing variants (rare and otherwise) into ancient datasets from modern reference panels, may benefit future studies using ancient genomic data to look at relationships between closely related groups in the past. The extraordinary level of aDNA preservation seen in the samples from Xaltocan (Mata-Miguez 2016) may facilitate these further analyses.

Conclusions

Overall, the ancient DNA analysis of Xaltocan has yielded substantial insights about the population dynamics in the town itself and about Central Mexico more broadly. Evidence from the uniparental data suggests that during the Early Period, there were two ethnically distinct neighborhoods. There is also evidence from the uniparental data for major genetic changes at the site following the Tepanec conquest and Aztec annexation of Xaltocan. However, the genome-wide data gathered thus far do not provide support for these conclusions. Both the uniparental

and genome-wide data support the hypothesis that close genetic relatives were buried together across time periods at Xaltocan, often in or near housing structures. The genome-wide data also show that the ancient residents of Xaltocan are closely related to many populations living and Central and Southern Mexico today. However, it does not appear that the ancient Xaltocan individuals share significantly more genetic affinity with modern residents of the three Xaltocans. Future work collecting higher coverage genomic data from the ancient residents of Xaltocan may help to disentangle the close relationships we see between time periods at the site and help resolve the discordant results between the uniparental and genome-wide datasets. Ongoing paleogenomic work at Xaltocan and other sites in Central Mexico will further add to our understanding of the rise of pre-Aztec city states, the consolidation and expansion of the Aztec empire, and the effects of early European colonialism in this region.

Methods

Samples

Skeletal remains from 42 individuals were recovered from six houses (Operation G, Operation Y, Zocalo A, Zocalo C, Structure 122, and Structure 124; Figure S1) during archaeological research at Xaltocan (Brumfiel 2005a; Brumfiel 2007; Brumfiel and Rodríguez-Alegría 2010; Overholtzer 2012; Overholtzer 2013; DeLucia and Overholtzer 2014). Operation G, Zocalo A, and Zocalo C were multi-room houses in the interior of the ancient island where individuals were buried

under house floors, whereas Operation Y, Structure 122, and Structure 124 were single-room houses on the periphery of the island where individuals were buried in outside patios. Archaeological excavations at Xaltocan were carried out with support and appropriate permissions from the Mexican National Institute for Anthropology and History, the town's delegates, and local property owners.

DNA Extraction and Initial QC

Ancient DNA was extracted from 38 of the skeletal samples following published protocols, using strict procedures to prevent and detect contamination (Rohland and Hofreiter 2007; Bolnick et al. 2012). 26 of these individuals date to the Early Period (N=8 interior; N=18 periphery), 8 individuals date to the Tepanec Period, 3 individuals date to the Aztec period, and 5 individuals who could not be assigned to the Tepanec or Aztec period based on archaeological evidence. Each extract was screened for aDNA preservation and authenticity using a combination of mitochondrial DNA from the first hypervariable region, 23 Y-chromosome short tandem repeats (STRs) using the PowerPlex® 23 System kit (Promega), and 15 autosomal STRs using the PowerPlex® 16 System kit (Promega). Samples with successful and replicable recovery of mtDNA and several of the STR markers were promoted to genome-wide analysis. Genetic sex for each sample was assigned by targeting a length dimorphism in the amelogenin gene on the X and Y chromosomes. All DNA extraction and initial QC was done by Dr. Jaime Mata-Míguez at the University of Texas at Austin.

Library Construction and Genomic QC

Nineteen of the 38 extracted individuals were selected for genome-wide analysis based on the results of our initial QC. Double-stranded libraries with truncated Illumina adapters were prepared using partial UDG treatment following previously published protocols (Rohland et al. 2015; Haak et al. 2015). All libraries were prepared by Dr. Jaime Mata-Míguez at the University of Texas at Austin. Libraries were then shipped to the lab of Dr. David Reich at Harvard University for genomic capture, sequencing, and QC. There, in-solution enrichment for a targeted set of 1,237,207 SNPs (1240K capture) was performed on each library (Mathieson et al. 2015) and sequenced on an Illumina NextSeq500 instrument. Read pairs were merged with the expected barcodes that overlapped by at least 15 bases, mapped the merged sequences to the human reference genome hg19 using the 'samse' command in bwa v.0.6.1 (Li and Durbin 2009) and then duplicated sequences were removed. DNA authenticity was evaluated by estimating the rate of mismatching to the consensus mitochondrial sequence, and also by requiring that the rate of damage at the terminal nucleotide was at least 3% (Fu et al. 2013; Sawyer et al. 2012). DNA authenticity was further evaluated by looking at the ratio of X-to-Y chromosome reads and estimating X-chromosome contamination in males based on the rate of heterozygosity (Korneliussen et al. 2014). Results of this genomic capture, along with QC metrics for each sample can be found in **Table 1.4**. Pairs of related ancient individuals were identified based on the proportion of sites covered

Table 1.4 – Ancient Xaltocan sample information

SampleID 1	SampleID 2	Skeletal element	Date	Label	Sex	mtDNA
I2834	ZA.1	long bone	1240-1395 CE	Early Interior	M	A2g
I6508	G.1	long bone	1240-1395 CE	Early Interior	F	A2
I6509	ZC.1	long bone	1240-1395 CE	Early Interior	F	A2x
I2838	E14.5	long bone	1240-1395 CE	Early Periphery	F	D1
I6969	Y2.7	tooth	1200-1500 CE	Early Periphery	F	B2
I2131	Y3.8.1	tooth	1240-1350 CE	Early Periphery	F	A2
I2835	Y3.10A	tooth	1240-1395 CE	Early Periphery	F	D1
I2836	Y3.4	long bone	1240-1395 CE	Early Periphery	F	B2c2b
I2839	Y3.6	tooth	1240-1395 CE	Early Periphery	F	B2
I2844	E10.1	tooth	1395-1521 CE	Tepanec	M	A2
I6966	E.8.2	tooth	1400-1470 CE	Tepanec	F	A2u
I6507	E8.1	tooth	1390-1460 CE	Tepanec	M	A2j
I2842	E7.1	tooth	1400-1450 CE	Tepanec	M	A2u
I2843	E8.5	tooth	1390-1440 CE	Tepanec	F	A2u
I2841	E6.1	tooth	1410-1450 CE	Aztec	M	B2
I2846	E14.6	tooth	1430-1500 CE	Aztec	F	B2
I2845	E14.1	tooth	1395-1521 CE	Late Unknown	F	B2
I2847	E30.3	tooth	1395-1521 CE	Late Unknown	M	C1b3
I2848	E34.1	tooth	1330-1430 CE	Late Unknown	M	D1

in pairs of ancient samples from the same population that had identical allele calls using PLINK (Chang et al. 2015).

Merging Ancient Genomic Data with Modern Comparative Datasets

I combined the ancient genomic data with a comparative dataset of individuals from global populations, including Native American individuals who were masked for non-indigenous ancestry, for further analysis. The modern dataset was made up individuals genotyped on three different arrays. Individuals with >10% and SNPs with >5% missingness were removed from the dataset using PLINK v1.9 (Chang et al. 2015). Then I used the program *smartrel* from the EIGENSOFT package (Patterson et al. 2006) to identify and remove individuals from the dataset that were second-degree relatives or closer. The remaining individuals were phased using SHAPEIT2 (Delaneau et al. 2012). I next used RFMIX (Maples et al. 2013) to assign each chromosomal segment to its most likely ancestry for each Native American individual in the modern dataset. I used 30 YRI individuals, 30 CEU individuals, and 30 unadmixed PEL individuals from the 1000 Genomes Phase 3 dataset (1000 Genomes Project Consortium 2015) to represent the possible ancestral populations for this local ancestry assignment. After ancestry assignment, I removed SNPs from each individual with either a non-indigenous allele or a low-confidence (<90%) ancestry assignment. SNPs with one indigenous and one non-indigenous allele were also removed. I then merged all of the arrays with the ancient

individuals for a final dataset of 1045 individuals and 101,679 SNPs. Detailed information about the comparative dataset can be found in **Table A2**.

ADMIXTURE Analysis

I performed model-based clustering analysis using ADMIXTURE (Alexander et al. 2009) on the combined dataset. First I pruned the dataset by linkage disequilibrium using PLINK (Chang et al. 2015) with the flag –indep-pairwise 200 1 0.4, leaving 66080 SNPs. I ran admixture with the cross validation flag from K=2 to K-20 clusters, with 10 replicates for each value of K (**Figure 1.16)**. For each K value, the replicate with the highest log likelihood was kept.

Principal Component Analysis and Discriminant Function Analysis

I performed principal component analysis of the pruned combined dataset using the *smartpca* program in EIGENSOFT (Patterson et al. 2006). I computed principal components on all of the modern individuals in the dataset and projected the ancient individuals using the *lsqproject* option. I used a broken stick model in R to determine the number of PC axes to retain for discriminant function analysis. I then used a MANOVA to test whether the ancient groups were significantly different, and a linear discriminant analysis to visualize these differences.

f*-statistics and *qpWave

I computed *f*-statistics on the combined dataset using ADMIXTOOLS (Patterson et al. 2012). I used *qpDstat* with *f4mode* for *f*4-statistics and *qp3Pop* for outgroup *f*3-statistics. I computed standard errors using a weighted block jackknife

over 5-Mb blocks. I also attempted to model various modern groups as a mixture of a number of ancient and modern American populations using the *qpWave* tool in ADMIXTOOLS (Haak et al. 2015).

Supplemental Figures

Figure 1.14 – ADMIXTURE plots at K=2-20

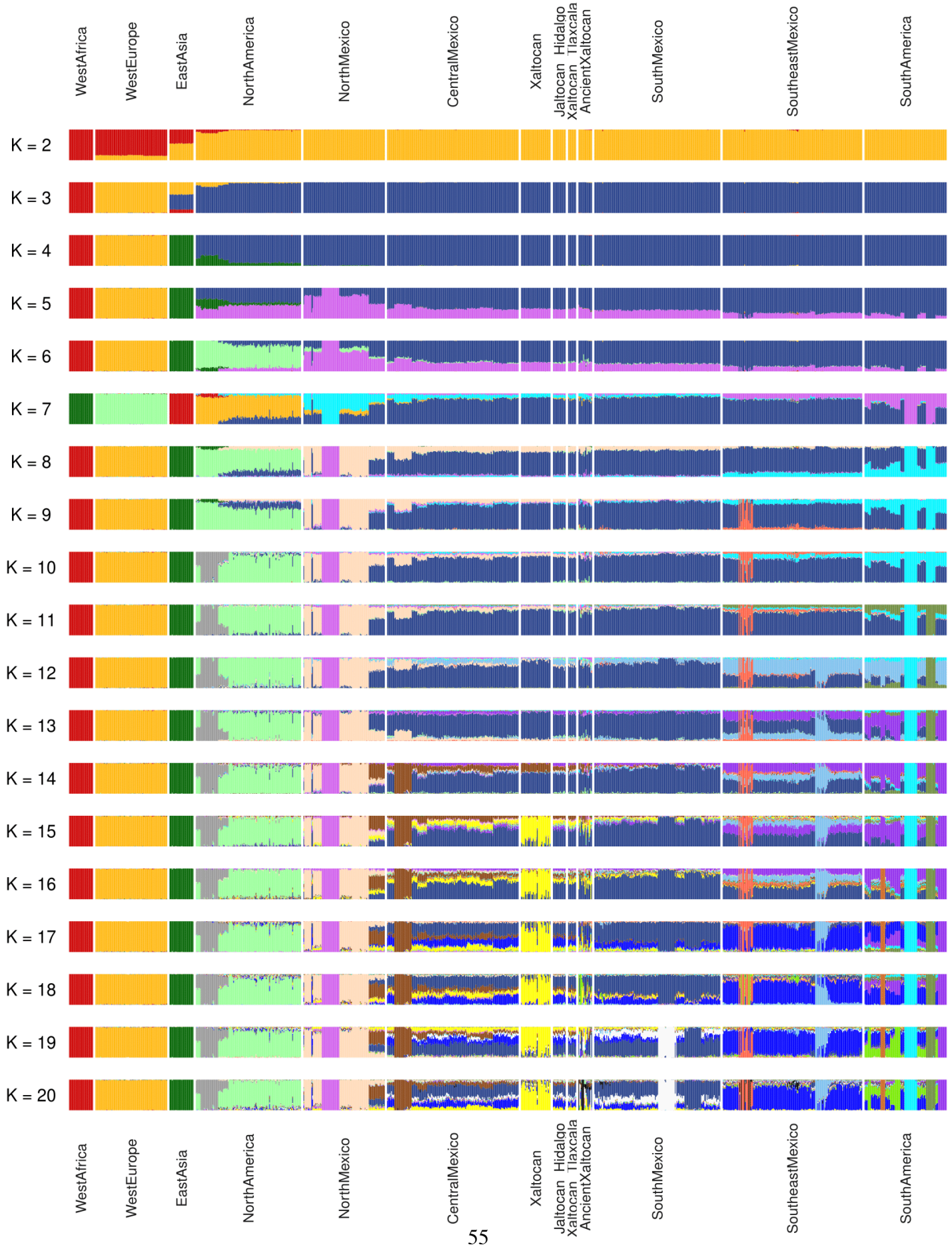


Figure 1.15 – ADMIXTURE proportions for individual ancient samples at K=15

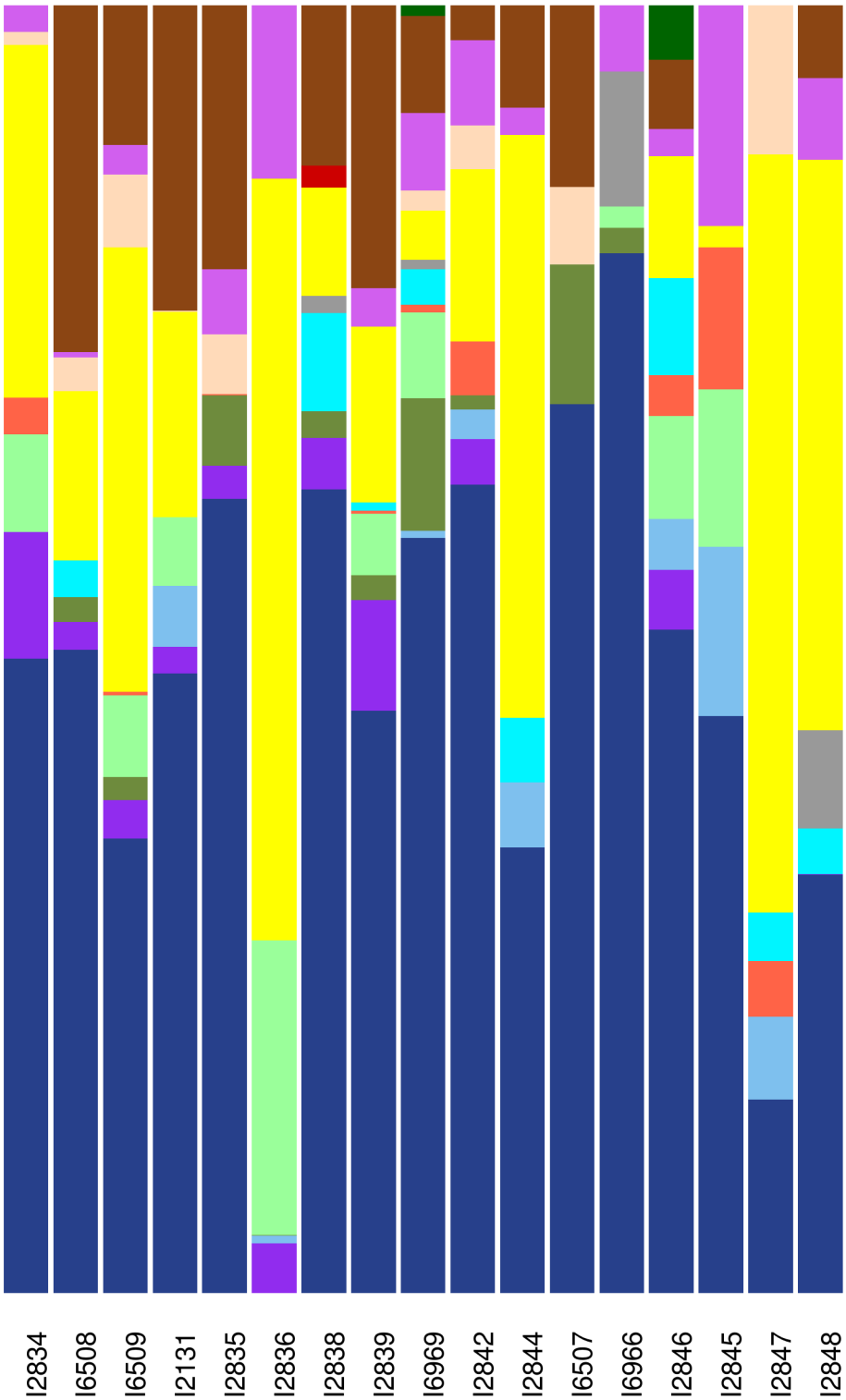
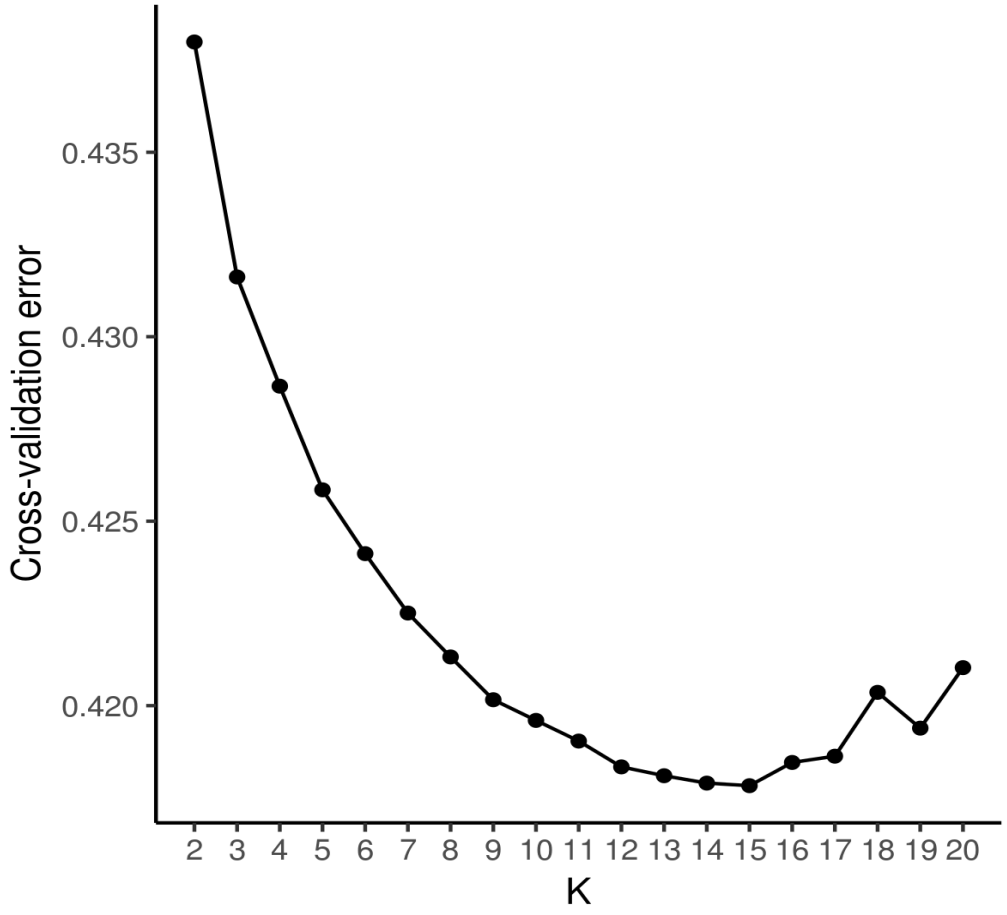


Figure 1.16 – Cross-validation errors by K. K=15 has the lowest cross-validation error.



Chapter 2: Exploring Native American Genetic Structure in Mexico and the Effects of European Colonialism.

Introduction

While recent studies have shed light on the initial peopling of the Americas using genomic data (Rasmussen et al. 2015, 2014; Raghavan et al. 2015; Reich et al. 2012; Skoglund et al. 2015), less work has been done applying genomic datasets to questions about the more recent past. One area of particular interest is the genetic impact of European colonialism in the Americas. The genetic work that has been done on this topic has shown that the genetic impact of these historic events varies widely across the American continents. Brazil and many Caribbean island populations, for example, show much higher proportions of African ancestry than other populations in the Americas (Kehdy et al. 2015; Moreno-Estrada et al 2013). The populations of Argentina and Uruguay, on the other hand, show comparatively higher levels of European ancestry (Homburger et al. 2015; Salzano and Sans 2014). This study focuses on the patterns seen in population genomic data from Mexico, an ideal place to study both variation in indigenous American genetic diversity and the genetic impacts of European colonialism in the Americas.

Mexico constitutes a unique study area within the Americas because the country is home to large numbers of indigenous people. More than 16 million self-identified indigenous individuals (~15% of the population) live in Mexico today (Hansen et al. 2017), organized into 68 federally recognized ethnic groups

throughout Mexico. Because of these large numbers, working with these groups makes it possible to better reconstruct genetic diversity prior to European colonization than in other places in the Americas and to understand the genetic history of colonialism in the country. Since the Spanish conquest of Mexico in 1521, Mexico has received a substantial number of immigrants from around the world (though primarily from Europe and Africa). For instance, about 20,000 Spanish individuals and 15,000 enslaved African individuals lived in Mexico by the mid-16th century, making up about 1.5% of the total population at the time (García-Martínez 2010). This influx of new people led to large amounts of admixture between these groups and the indigenous peoples already in the area, primarily in the urban areas where they settled (Moreno-Estrada 2014). However, the effects of European colonization outside of the urban colonial centers have been less well studied, and not recognized until recently (citation?).

Along with this influx of people from elsewhere came sweeping changes to societal organization within Mexico, but this may have been more tempered in rural areas. Throughout the colonial period, most indigenous people in Mexico lived in relatively small communities scattered across the country, with little Spanish influence (Socolow 1996). The Spanish created a system of encomiendas shortly after the conquest, in which native communities were subject to a Spanish landlord who was tasked with collecting tribute from indigenous groups and supposedly defending and Christianizing them. However, many communities remained largely

independent, with their own leaders retaining substantial political control (Gibson 1964;; Endfield 2008; Restall and Schwaller 2011). Spanish colonists referred to these communities as pueblos de indios (literally, "Indians' towns"; Restall and Schwaller, 2011), and interactions between Spanish authorities and the native peoples in such communities may have been very limited. In addition, some laws barred Spaniards and Africans from living in pueblos de indios (Restall and Schwaller 2011) or trading with native individuals (Gibson 1964), and marriages between Spaniards and native individuals were not customarily endorsed by the church (Curcio-Nagy 2011). This history suggests that the effects of admixture in rural indigenous communities are less than those seen in urban groups and likely occurred later in time.

As with the extent of admixture in rural communities, the effects of epidemics and severe demographic crashes in indigenous populations — recorded in colonial Spanish documents — remain unclear for more rural groups in Mexico. The spread of European-introduced diseases began shortly after Spanish conquistador Hernán Cortés landed near the present-day city of Veracruz in 1519 and began his military campaign against the Aztec empire (Acuna-Soto et al. 2002). In Central Mexico, the first documented epidemic was smallpox in 1519-20, as Cortés marched towards the Aztec capital city of Tenochtitlan, located in present-day Mexico City in Central Mexico. This initial epidemic was followed by several subsequent smallpox outbreaks, particularly in the late 16th and early 17th

centuries (Acuna-Soto et al. 2002). Historic records suggest a reduction in population size of >90% in Mexico overall (Ubelaker 2006). This large decrease in population is somewhat supported by previously published genetic data, although estimating the precise parameters of the bottleneck that different groups experienced during colonial times has proven difficult (Moreno-Estrada et al. 2014). It is also unclear from the historical records how far the smallpox epidemics, along with outbreaks of other infectious diseases, spread beyond the urban centers, or what genetic impacts they had.

This study uses a large dataset of genome-wide data gathered from populations around Mexico to address some of these pressing questions about the effects of European colonialism across the country and to better understanding of the indigenous genetic diversity within Mexico. This large dataset includes newly collected genomic data from 96 individuals from three towns in Central Mexico. These samples, collected originally as a comparative dataset for the ancient DNA analysis of the archaeological site of Xaltocan, are from the rural town of Xaltocan in the State of Mexico, as well as two towns named Xaltocan and Jaltocan in the nearby states of Tlaxcala and Hidalgo.

The town of Xaltocan in the State of Mexico was founded in ~900 AD by indigenous Otomi speaking peoples, and rose to power as a sizeable city-state between the 10th and 13th centuries (Brumfiel 2005a). In the mid-13th century, Xaltocan entered a war with several neighboring city-states, which it eventually lost,

significantly reducing its power in the region (Bierhorst 1992). Historical accounts claim that the original Otomi-speaking inhabitants of Xaltocan abandoned the town at this time, leaving it uninhabited for 40 years before being annexed into the Aztec empire, which sent new people to inhabit the town (Ixtlilxochitl 1975, 1977; Carrasco-Pizana 1987; Brumfiel 2005b; Hicks 1994). Oral histories collected by Dr. Jaime Mata-Míguez in the summer of 2015 suggested a possible connection between the ancient residents of Xaltocan and the ancestors of the people living in two other towns in Central Mexico (Xaltocan, Tlaxcala and Jaltocan, Hidalgo). He subsequently collected genetic samples from residents of these two towns to explore the genetic relationship among the three towns that share a name.

This newly collected data will also allow us to better assess the history of admixture at Xaltocan in the State of Mexico, as historical and archaeological evidence provide somewhat conflicting accounts about the extent of admixture in this community during the colonial period. Archaeological findings show that the native residents of Xaltocan adopted Spanish material culture, such Spanish dress, weaponry, and ceramics, likely to improve their social status (Rodríguez-Alegría, 2010). However, colonial documents suggest little interaction between the Spanish and the native residents of Xaltocan. In a 1569 letter addressed to the archbishop of Mexico, for instance, a Spanish priest mentioned that there were about 9,000 native individuals but only two married Spaniards who owned land in his parish, which included Xaltocan (Montúfar, 1897). Therefore, genetic data provides us an

excellent tool to see if the material culture found at Xaltocan during the colonial period was accompanied by substantial genetic inputs from the Spanish. Combining new genomic data with a detailed understanding of the history and archaeology of this location should provide important insights.

These large genomic datasets, along with the detailed knowledge of the history and archaeology of Xaltocan, allows me to test hypotheses on both a local and regional scale. Specifically, I will be testing the following hypotheses:

1. Admixture and non-indigenous genetic ancestry will be higher in Central Mexico, where the Spanish first arrived in Mexico and where their centers of power were located.
2. The estimated time of admixture events will be older in Central Mexico than in other regions of the country, reflecting colonial history.
3. Indigenous genetic diversity in Mexico will be geographically structured, and gene flow between groups will have been affected by the physical geography of the landscape they inhabit.
4. The genome-wide data analyzed here will show that the Xaltocan population experienced a genetic bottleneck following European colonization that can be detected with these genome-wide data.
5. The modern residents of Xaltocan will be closer genetically to the residents of Xaltocan, Tlaxcala and Jaltocan, Hidalgo, in line with the oral histories collected from these towns.

Results

Using the Affymetrix Axiom Human Origins Array, I collected genome-wide genotype data on ~620K SNPs from 47 current residents of the town of Xaltocan in the State of Mexico, as well as from 46 current residents of two other towns, Xaltocan in the Mexican state of Tlaxcala (XaltocanTlaxcala) and Jaltocan in the Mexican state of Hidalgo (JaltocanHidalgo). These data were combined with 1463 samples from urban Mestizo and indigenous Mexican groups as well as European and African populations to test my hypotheses regarding the genetic diversity and admixture history of Mexico. **Figure 2.1** shows a map of sampling locations in Mexico, and detailed information about the samples can be found in **Table B1**. For analyses of indigenous genetic structure, regions of the genome derived from recent African and European ancestors were identified using RFMIX (Maples et al. 2013) and removed.

Admixture History of Mexico

To better understand the extent of recent European and African admixture in Mexican populations, I first constructed a PCA using the unmasked genotype data. A PCA with the global comparative populations shows that the urban Mestizo populations skew more towards the European and African clusters than the indigenous Mexican groups (**Figure 2.2**). Highlighting each of the Xaltocans individually (**Figures 2.3-4**) shows the XaltocanTlaxcala population skewing more

Figure 2.1 – Map of sampling locations within Mexico. The locations of the Xaltocans are highlighted in red. Locations with the prefix "Cos_" show the cities where cosmopolitan Mestizo samples were collected.

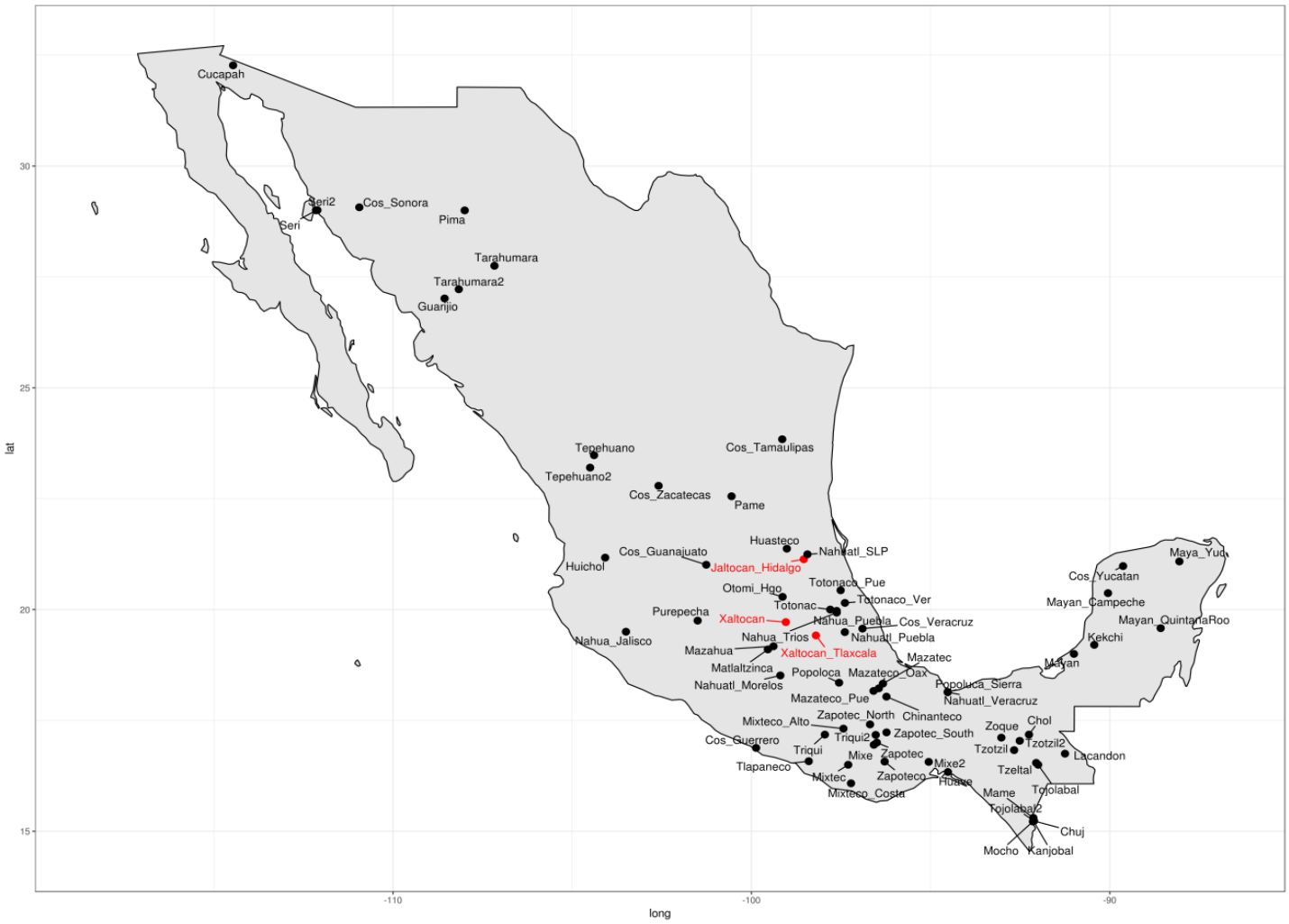


Figure 2.2 – Global PCA. Nat_Mexico = indigenous Mexicans, Cos_Mexico = cosmopolitan Mexican cohorts.

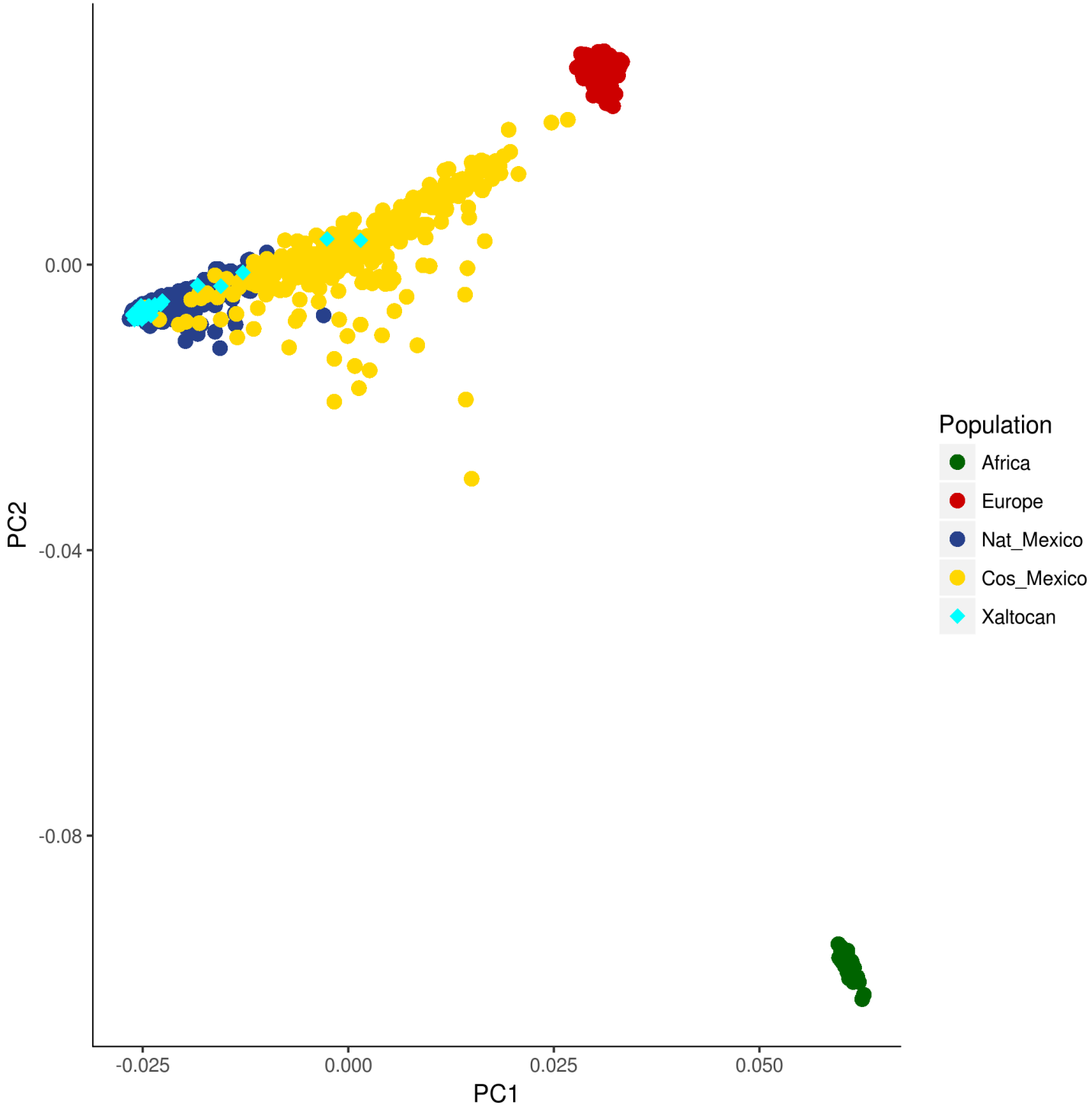


Figure 2.3 – PCA plot with global populations highlighting JaltocanHidalgo.

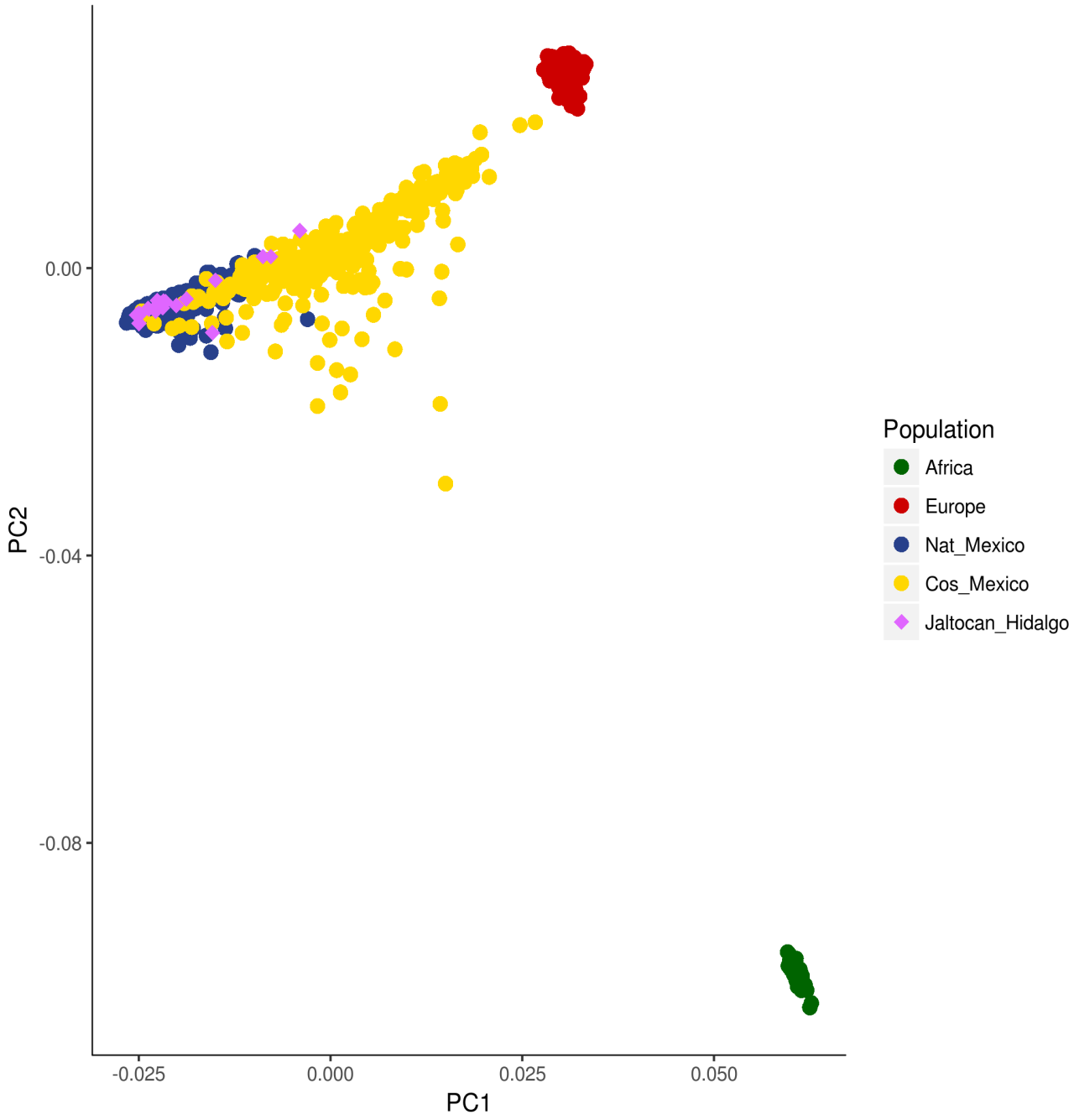
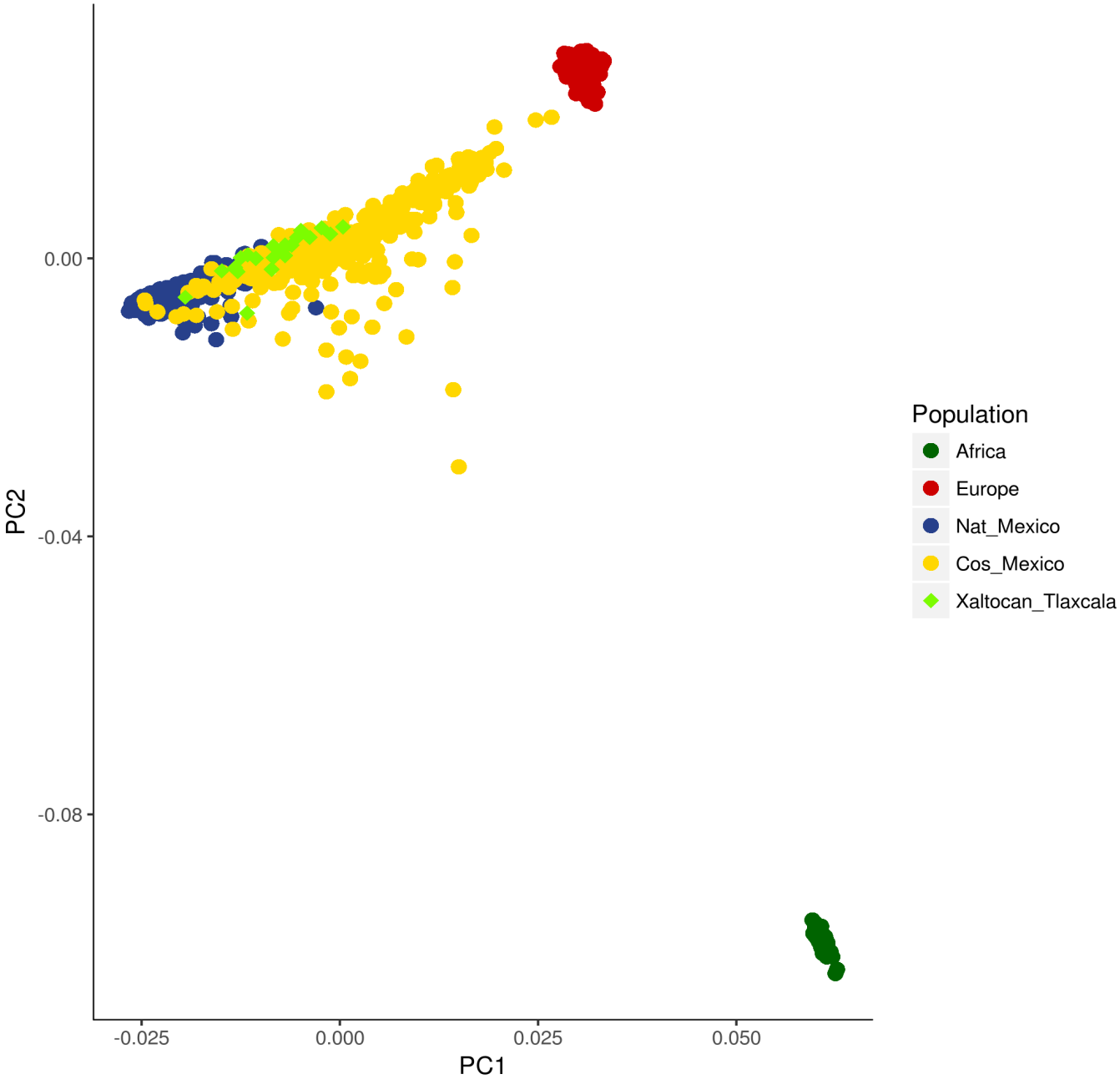


Figure 2.4 – PCA plot with global populations highlighting XaltocanTlaxcala.



towards Europeans relative to Xaltocan and JaltocanHidalgo, suggesting higher levels of recent European ancestry in that population.

ADMIXTURE analysis (Alexander et al. 2009) using the unmasked genotype data at K=3 (**Figure 2.5**) shows that compared to urban Mestizo populations, indigenous and rural groups (including the Xaltocans) have much lower proportions of European and African genetic ancestry. **Table 2.1** shows the average percent of ancestry for each population. Xaltocan has the highest amount of indigenous genetic ancestry on average (~95%), followed by JaltocanHidalgo (~91%), and lastly XaltocanTlaxcala (~71%). Most of the non-indigenous genetic ancestry found in these individuals comes from European sources; however, XaltocanTlaxcala shows an appreciable but low level of recent African ancestry (~2%).

I next binned estimates of genetic ancestry in two ways: the four broad regions of Mexico (North, Central, South, Southeast), and the urban Mestizo cohorts versus the indigenous and rural groups (**Table 2.2**). Urban Mestizo groups in Mexico have substantially more African (~5%) and European (~40%) ancestry on average than indigenous groups (~0.4% and 3% respectively). There is also variation in admixture proportions within indigenous groups from the different regions of Mexico. The North and South regions have the least amount of African ancestry on average (~0.2%), while the Southeast region has significantly more African ancestry (~0.7%) than the other regions (**Table 2.3**). A similar pattern is seen in the amount of European ancestry across regions, with Central Mexico (~5%

Figure 2.5 – ADMIXTURE plot at K=3 illustrating the low level of African and European genetic ancestry in indigenous Mexican populations compared to cosmopolitan cohorts.

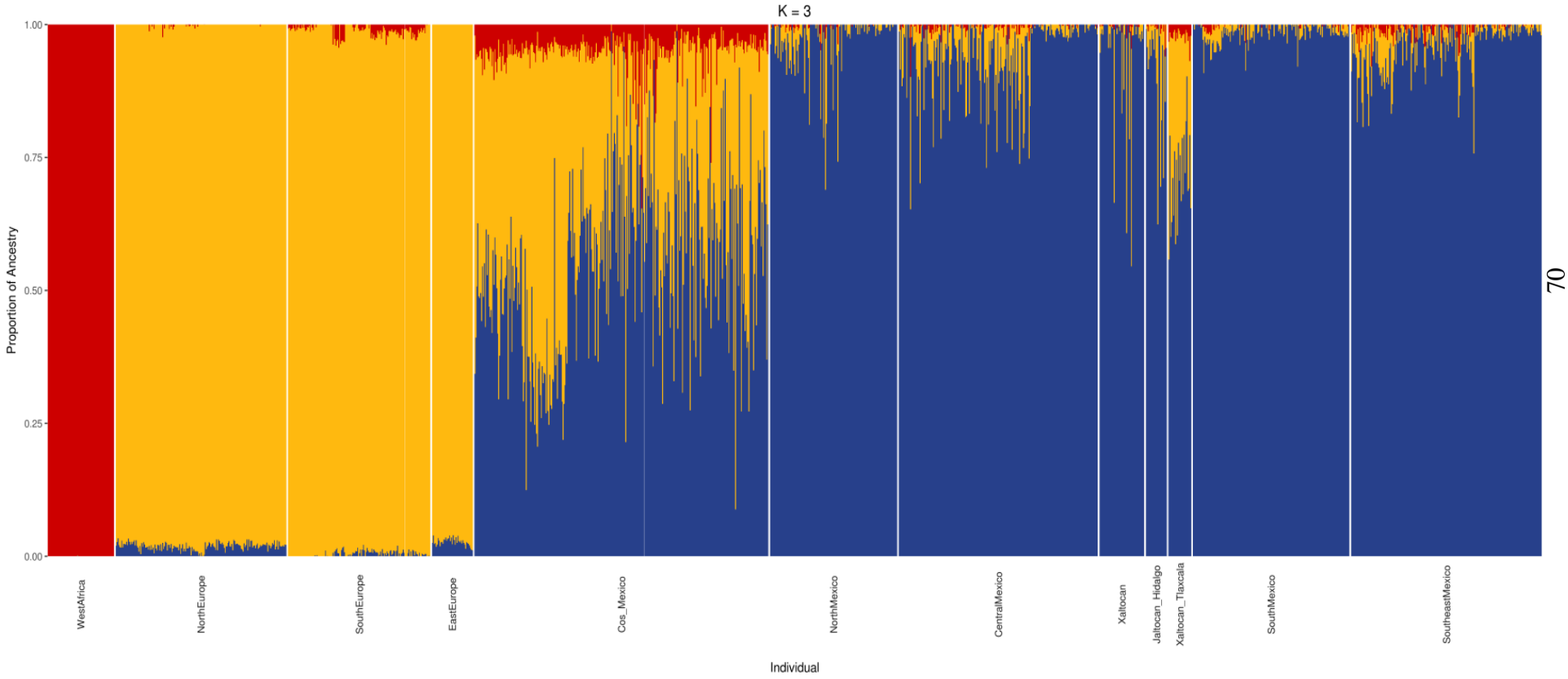


Table 2.1 – Genetic ancestry proportions inferred at K=3. NAM = Native American, EUR = European, AFR = African

Population	NAM	EUR	AFR
Xaltocan	0.954	0.042	0.004
JaltocanHidalgo	0.909	0.085	0.006
XaltocanTlaxcala	0.711	0.267	0.022

Table 2.2 – Admixture proportions by population. AFR= African, EUR= European, AMR= American

Population	Region	AFR	EUR	AMR	Population	Region	AFR	EUR	AMR
Cos_Mexico	Mexico	0.053	0.399	0.548	Chinanteco	South Mexico	0.002	0.011	0.987
Nat_Mexico	Mexico	0.004	0.032	0.964	Huave	South Mexico	0	0	1
Cos_Veracruz	Central Mexico	0.049	0.337	0.614	Mazatec	South Mexico	0.004	0.024	0.973
Cos_Guanajuato	Central Mexico	0.048	0.366	0.586	Mazateco_Oax	South Mexico	0.002	0.01	0.988
Cos_Zacatecas	North Mexico	0.058	0.442	0.5	Mazateco_Pue	South Mexico	0	0.003	0.997
Cos_Tamaulipas	North Mexico	0.053	0.43	0.517	Mixe	South Mexico	0	0.002	0.998
Cos_Sonora	North Mexico	0.046	0.611	0.343	Mixe2	South Mexico	0.001	0.02	0.979
Cos_Guerrero	South Mexico	0.074	0.256	0.67	Mixtec	South Mexico	0.013	0.031	0.956
Cos_Yucatan	Southeast Mexico	0.039	0.379	0.582	Mixteco_Alto	South Mexico	0.001	0.008	0.992
Matlaltzinca	Central Mexico	0.001	0.002	0.997	Mixteco_Costa	South Mexico	0.003	0.011	0.986
Mazahua	Central Mexico	0.002	0.005	0.992	Popoloca	South Mexico	0.001	0.014	0.984
Nahuatl_Morelos	Central Mexico	0.003	0.062	0.936	Tlapaneco	South Mexico	0	0	1
Otomi_Hgo	Central Mexico	0.001	0.015	0.984	Triqui	South Mexico	0	0.004	0.996
Xaltocan	Central Mexico	0.004	0.042	0.954	Triqui2	South Mexico	0	0.004	0.996
Nahua_Puebla	Central Mexico	0.005	0.13	0.864	Zapotec	South Mexico	0.004	0.039	0.957
Nahua_Trios	Central Mexico	0.004	0.002	0.994	Zapotec_North	South Mexico	0	0.013	0.987
Nahuatl_Puebla	Central Mexico	0.002	0.007	0.991	Zapotec_South	South Mexico	0.001	0.018	0.981
Totonac	Central Mexico	0.001	0.041	0.958	Zapoteco	South Mexico	0	0.011	0.989
Totonaco_Pue	Central Mexico	0	0.002	0.998	Chol	Southeast Mexico	0.001	0.011	0.988
Totonaco_Ver	Central Mexico	0	0.012	0.988	Chuj	Southeast Mexico	0	0.011	0.989
Xaltocan_Tlaxcala	Central Mexico	0.022	0.267	0.711	Kanjobal	Southeast Mexico	0	0.005	0.995
Huichol	Central Mexico	0.002	0.009	0.99	Kekchi	Southeast Mexico	0	0.013	0.987
Nahua_Jalisco	Central Mexico	0.019	0.1	0.882	Lacandon	Southeast Mexico	0.008	0.02	0.972

Table 2.2, cont. – Admixture proportions by population. AFR= African, EUR= European, AMR= American

Population	Region	AFR	EUR	AMR	Population	Region	AFR	EUR	AMR
Purepecha	Central Mexico	0.011	0.096	0.893	Mame	Southeast Mexico	0.008	0.016	0.976
Guarijio	North Mexico	0	0.002	0.998	Maya_Yuc	Southeast Mexico	0.024	0.065	0.911
Pame	North Mexico	0	0.008	0.992	Mayan	Southeast Mexico	0.001	0.015	0.984
Tarahumara	North Mexico	0.006	0.064	0.931	Mayan_Campeche	Southeast Mexico	0.015	0.069	0.916
Tarahumara2	North Mexico	0	0.007	0.993	MayanQuintanaRoo	Southeast Mexico	0.014	0.069	0.916
Tepehuano	North Mexico	0.005	0.044	0.951	Mocho	Southeast Mexico	0.002	0.024	0.974
Tepehuano2	North Mexico	0	0.021	0.979	Nahuatl_Veracruz	Southeast Mexico	0	0.001	0.999
Huasteco	Central Mexico	0	0.005	0.995	Popoluca_Sierra	Southeast Mexico	0.002	0.002	0.997
Jaltocan_Hidalgo	Central Mexico	0.007	0.084	0.909	Tojolabal	Southeast Mexico	0.007	0.031	0.963
Nahuatl_SLP	Central Mexico	0	0.022	0.978	Tojolabal2	Southeast Mexico	0.005	0.012	0.983
Cucapah	North Mexico	0	0	1	Tzeltal	Southeast Mexico	0	0.007	0.993
Pima	North Mexico	0.002	0.024	0.974	Tzotzil	Southeast Mexico	0	0.015	0.985
Seri	North Mexico	0.002	0.019	0.978	Tzotzil2	Southeast Mexico	0	0.006	0.994
Seri2	North Mexico	0	0	1	Zoque	Southeast Mexico	0.002	0.013	0.985

Table 2.3 –P Values for African ancestry between regions

Central Mexico	0.0001161		
South Mexico	0.5323	1.04E-06	
Southeast Mexico	3.19E-07	0.04778	8.06E-10
	North Mexico	Central Mexico	South Mexico

on average) and Southeast Mexico (~3%) having significantly more European ancestry (**Table 2.4**) than the North (~2%) and South (~1%) of the country.

To estimate the timing of admixture into each of the three Xaltocan populations and a number of other indigenous Mexican groups, I used the program ALDER (Loh et al. 2013). ALDER uses the weighted LD curves of two potential "source" populations and an admixed "recipient" population to estimate the timing and extent of the admixture event leading to the recipient population. To produce a LD curve, one calculates the admixture LD between two populations at each pair of markers. As the genetic distance between the markers increases, the admixture LD between them decays at an exponential rate, producing the LD curve (Moorjani et al. 2011). Power analysis has shown that as sample size for the recipient population decreases below 20 individuals, the error rate for estimates of admixture time increase substantially, particularly for older admixture events. For this reason and because the admixture events in Mexico are likely to be very recent, I chose to run ALDER only for populations with >10 individuals. For my analysis, I used the CEU (Central European Utah), YRI (Yoruba in Ibadan, Nigeria) and unadmixed PEL (Peruvians from Lima, Peru) individuals from the 1000 Genomes dataset as surrogates for the source populations contributing to the present-day Xaltocan populations.

Results of these statistics (**Table 2.5**) suggest that European and African gene flow entered each of the three Xaltocan populations nearly 8 generations ago,

Table 2.4 –P Values for European ancestry between regions

Central Mexico	2.10E-06		
South Mexico	0.7541	1.24E-08	
Southeast Mexico	7.24E-06	0.29	4.83E-09
	North Mexico	Central Mexico	South Mexico

Table 2.5 – Alder results for the Xaltocans. CEU = Central European Utah; YRI = Yoruba Ibadan

Test_Pop	Ref_Pop	P_value	Z_score	Estimated Admixture Time (generations)
Xaltocan	CEU	1.80E-09	8.11	7.81 +/- 0.96
Xaltocan, Tlaxcala	CEU	1.70E-16	14.2	7.62 +/- 0.54
Xaltocan, Tlaxcala	YRI	1.70E-16	11.66	7.64 +/- 0.65
Jaltocan, Hidalgo	CEU	0.00011	9.1	7.47 +/- 0.82

in the late 18th or early 19th centuries. European and African gene flow occurred at about the same time in groups across Mexico (between ~3.4-15.3 and ~5.4-18.9 generations ago respectively). Comparing regions within Mexico (**Table 2.6**), South Mexico has the oldest estimated admixture times (average ~11.3 generations), followed by Central Mexico (~8.5 generations).

Indigenous Genetic Variation in Mexico

I used the genotyped data masked to remove regions of recent European and African ancestry to better understand the indigenous genetic variation within Mexico. The PCA constructed using these masked genotypes depicts the geographic relationships among the indigenous Mexican groups. **Figure 2.6** shows that one North Mexican group, the Seri, is an outlier in PC2. This has been reported elsewhere (Moreno-Estrada et al. 2014), and when the Seri are removed from the analysis, one can see a clear northwest to southeast cline in the PCA, reflecting the geographic relationship among populations in Mexico (**Figure 2.7**). Xaltocan clusters with the other populations from Central Mexico, although this population appears somewhat distinct. JaltocanHidalgo and XaltocanTlaxcala, however, cluster firmly in the middle of the Central and South Mexican groups.

An ADMIXTURE (Alexander et al. 2009) analysis at K=11, which had the lowest cross-validation error, shows the variation in Mexico very clearly, with several components exhibiting the northwest-to-southeast clines seen in previous analyses (Moreno-Estrada et al. 2014). A few isolated populations (including the

Table 2.6 – Estimated admixture times. CEU = Central European Utah; YRI = Yoruba Ibadan

Test_Pop	Region	Ref_Pop	Admixture time	CI
Pima	North Mexico	CEU	6.91	+/- 1.41
		YRI	7	+/- 1.34
Seri	North Mexico	YRI	6.36	+/- 1.82
		CEU	4.3	+/- 0.93
Tarahumara	North Mexico	CEU	7.59	+/- 0.78
		YRI	7.87	+/- 0.98
Tepehuano	North Mexico	YRI	7.61	+/- 0.73
		CEU	6.99	+/- 0.79
Huichol	Central Mexico	YRI	16.36	+/- 2.60
		CEU	13.71	+/- 1.61
NahuaJalisco	Central Mexico	CEU	6.61	+/- 0.80
		YRI	7.49	+/- 0.95
NahuaPuebla	Central Mexico	CEU	6.27	+/- 0.58
		YRI	6.44	+/- 0.72
NahuaTrios	Central Mexico	YRI	8.12	+/- 0.82
		CEU	7.63	+/- 0.79
Purepecha	Central Mexico	CEU	7.97	+/- 0.69
		YRI	8.75	+/- 0.75
Totonac	Central Mexico	YRI	8.29	+/- 1.11
		CEU	7.85	+/- 1.10
Xaltocan	Central Mexico	CEU	7.81	+/- 0.96
		YRI	9.67	+/- 1.01
XaltocanTlaxcala	Central Mexico	CEU	7.62	+/- 0.54
		YRI	7.64	+/- 0.65
JaltocanHidalgo	Central Mexico	CEU	7.47	+/- 0.82
		YRI	8.03	+/- 0.79
Triqui	South Mexico	YRI	14.19	+/- 3.33
		CEU	10.53	+/- 1.63
ZapotecNorth	South Mexico	YRI	12.21	+/- 1.74
		CEU	11.26	+/- 1.77
ZapotecSouth	South Mexico	CEU	9.89	+/- 1.33
		YRI	9.98	+/- 1.24
Lacandon	Southeast Mexico	YRI	6.16	+/- 0.72
		CEU	6.5	+/- 0.80
Mayan	Southeast Mexico	CEU	8.69	+/- 1.09
		YRI	10.02	+/- 1.05
MayanCampeche	Southeast Mexico	CEU	6.53	+/- 0.64
		YRI	6.53	+/- 0.40
MayanQuintanaRoo	Southeast Mexico	CEU	10.32	+/- 0.78
		YRI	9.92	+/- 0.99
Tojolabal	Southeast Mexico	YRI	9.18	+/- 1.85
		CEU	8.14	+/- 1.13
Tzotzil	Southeast Mexico	YRI	8.48	+/- 1.73
		CEU	6.46	+/- 1.24

Figure 2.6 – PCA of indigenous American ancestry components in Mexican populations

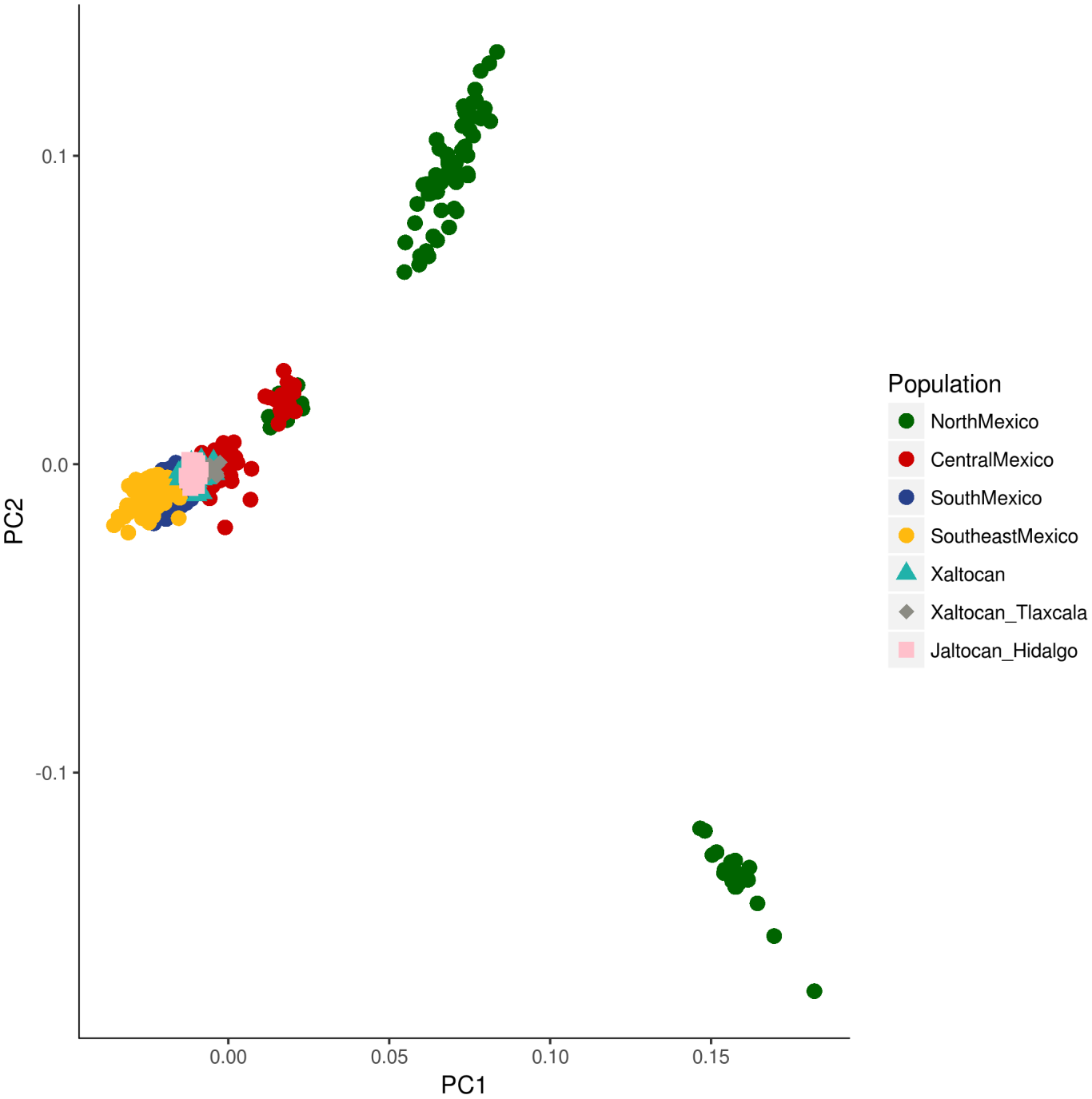
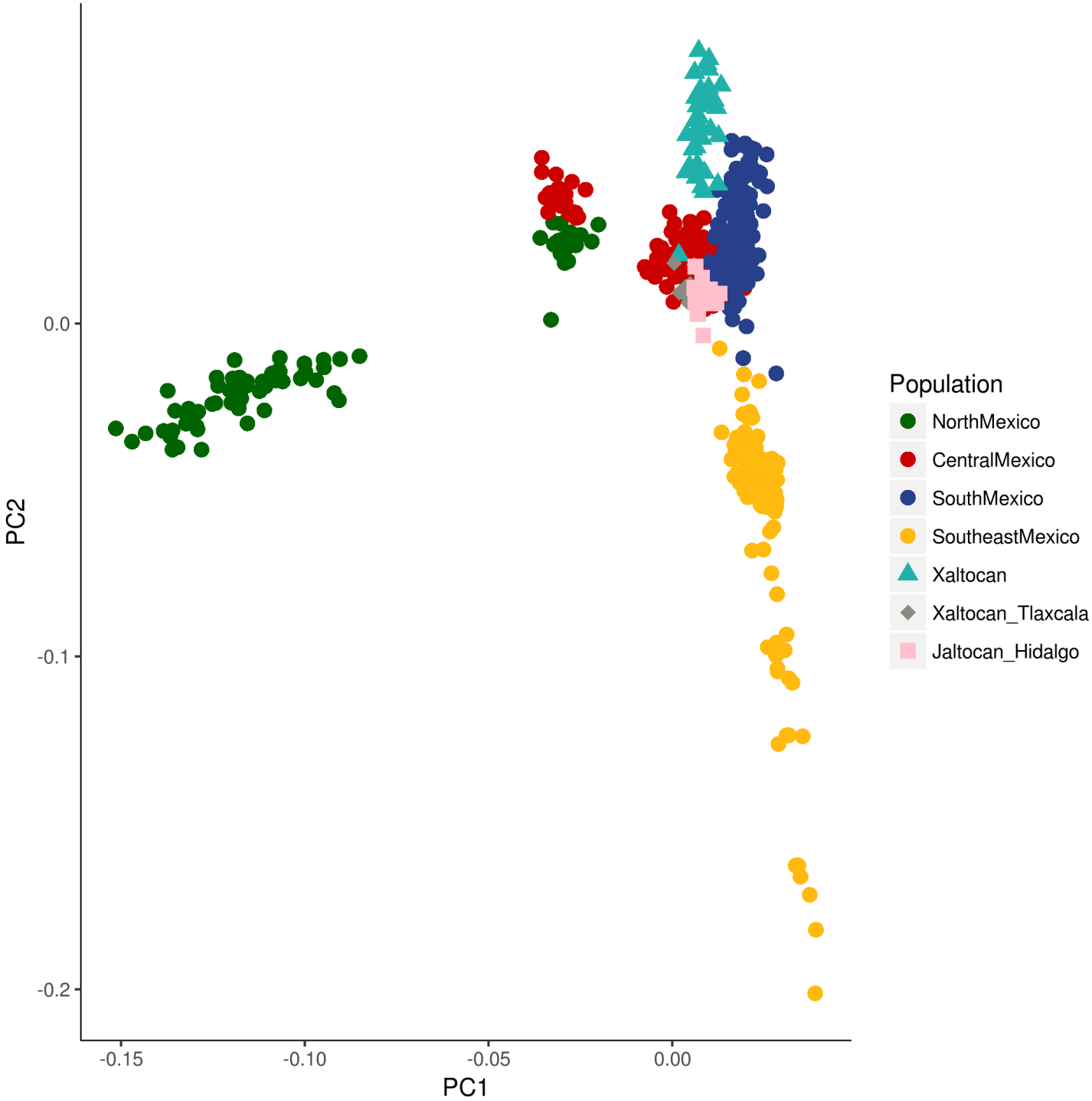


Figure 2.7 – PCA of indigenous American ancestry components in Mexican populations (without the Seri)



Seri in the Northwest, and the Lancandon and Maya from the state of Quintana Roo in the Southeast, as well as Xaltocan), separate into their own components (**Figure 2.8**), likely indicative of a large amount of drift between these populations and other groups in the analysis.

To further explore the genetic affinities of the three Xaltocan populations with others across Mexico, I calculated outgroup-*f*3 statistics of the form (X,Y;Yoruba, where X is a Xaltocan group and Y is a population from the comparative dataset) using the masked genotype data. The value of the outgroup-*f*3 statistic is proportional to the relatedness between PopulationX and PopulationY, and therefore, greater values mean that PopulationX and PopulationY share a higher degree of genetic drift. Outgroup-*f*3 results show qualitatively similar results between the three Xaltocan populations. All are quite similar to various groups in Central and South Mexico, particularly the NahuatlMorelos, and genetically distant to population from North Mexico and, to a lesser extent, those in the Southeast (**Figures 2.9-10,11**).

I also used *f*4 statistics of the form (YRI,X;Y,Z, where X is one of the Xaltocan groups and Y and Z are every pairwise combination of the comparative populations plus the ancient groups) to test for significant relationships among the Xaltocans and between each Xaltocan and all of the modern groups in my comparative dataset. This test evaluates if the data are consistent with a four-population tree of the form: (Outgroup, Population X; PopulationY, PopulationZ). The value of the statistic under

Figure 2.8 – ADMIXTURE results at K=11.

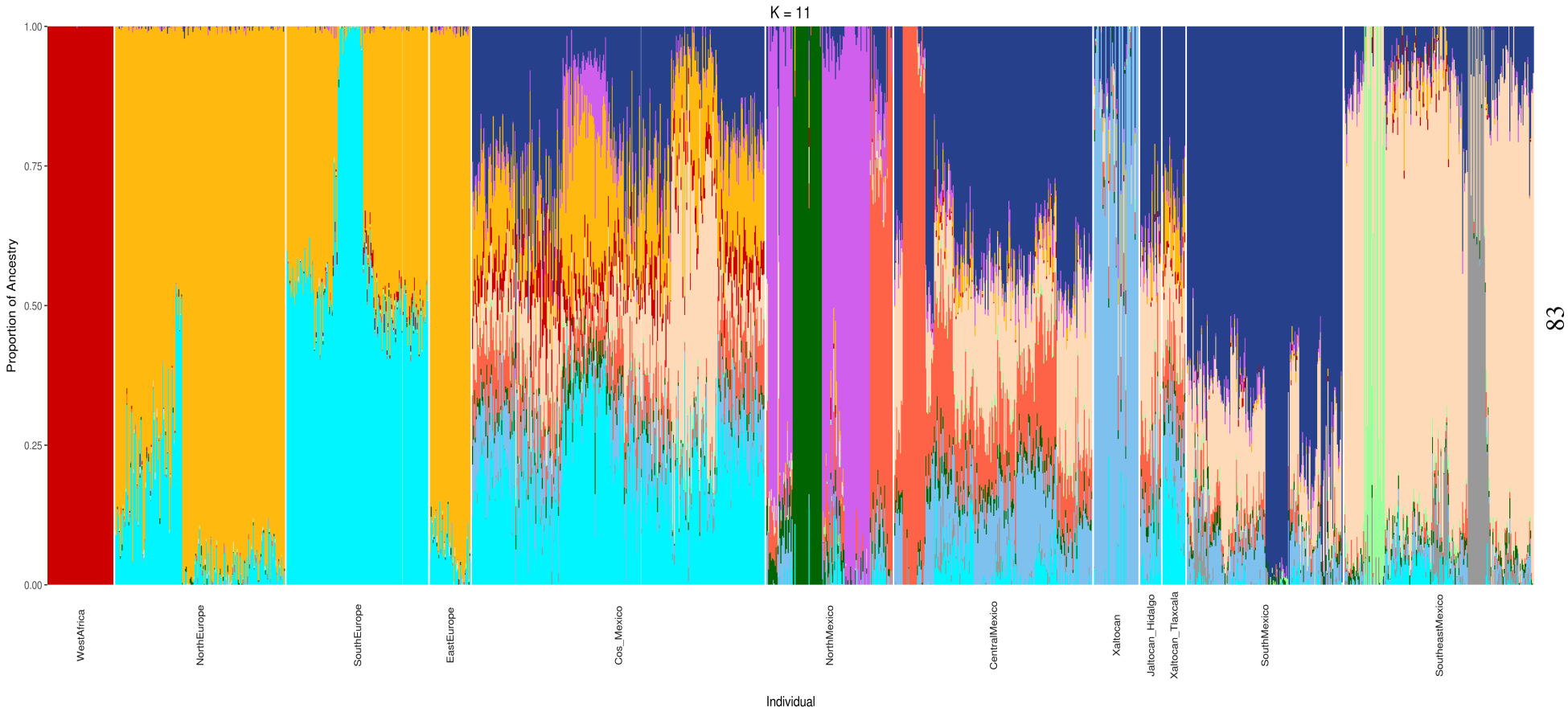


Figure 2.9 – Map of outgroup-f3 results for Xaltocan. Warmer colors show a closer relationship to the population of interest.

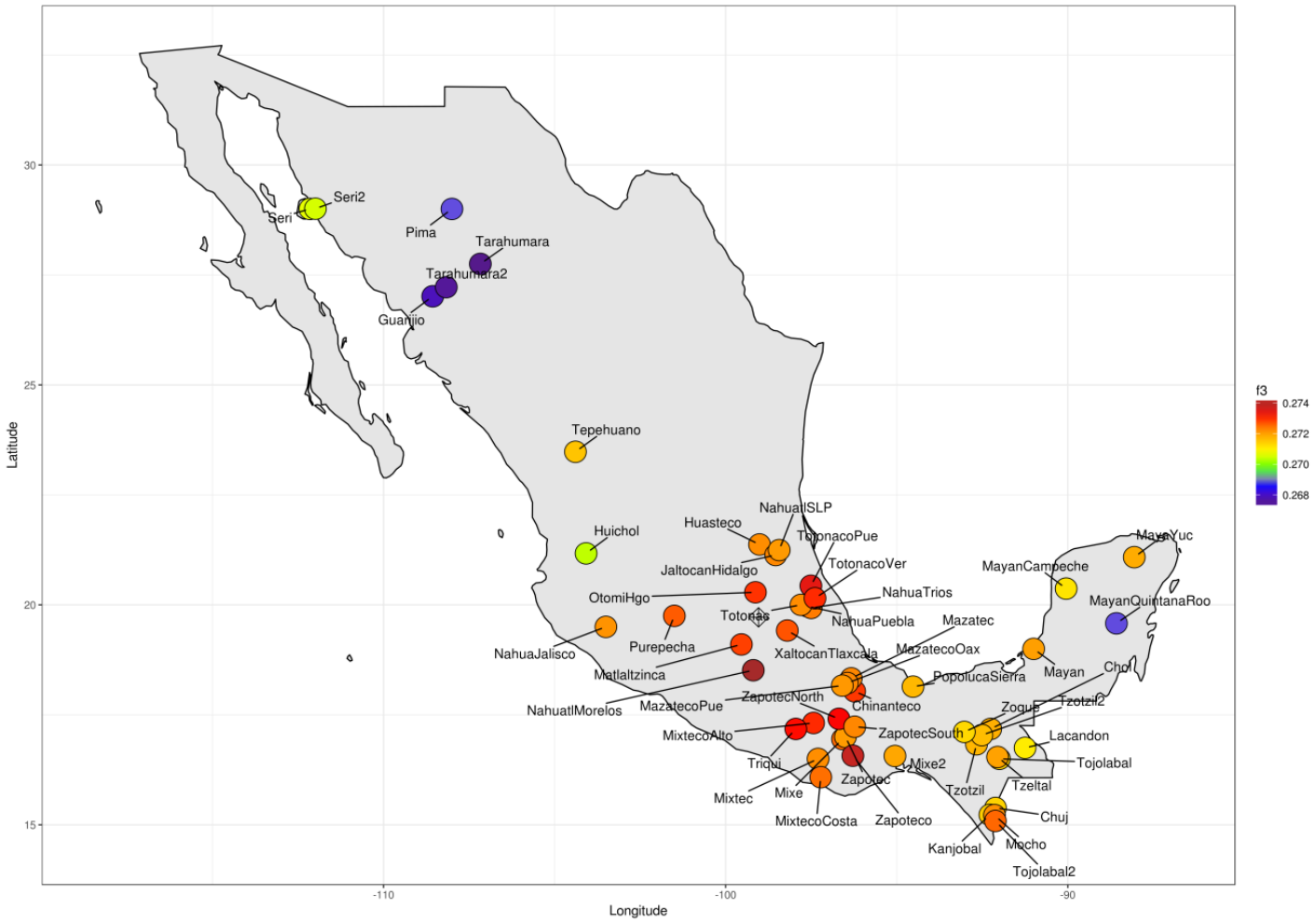


Figure 2.10 – Map of outgroup-f3 results for JaltocanHidalgo. Warmer colors show a closer relationship to the population of interest.

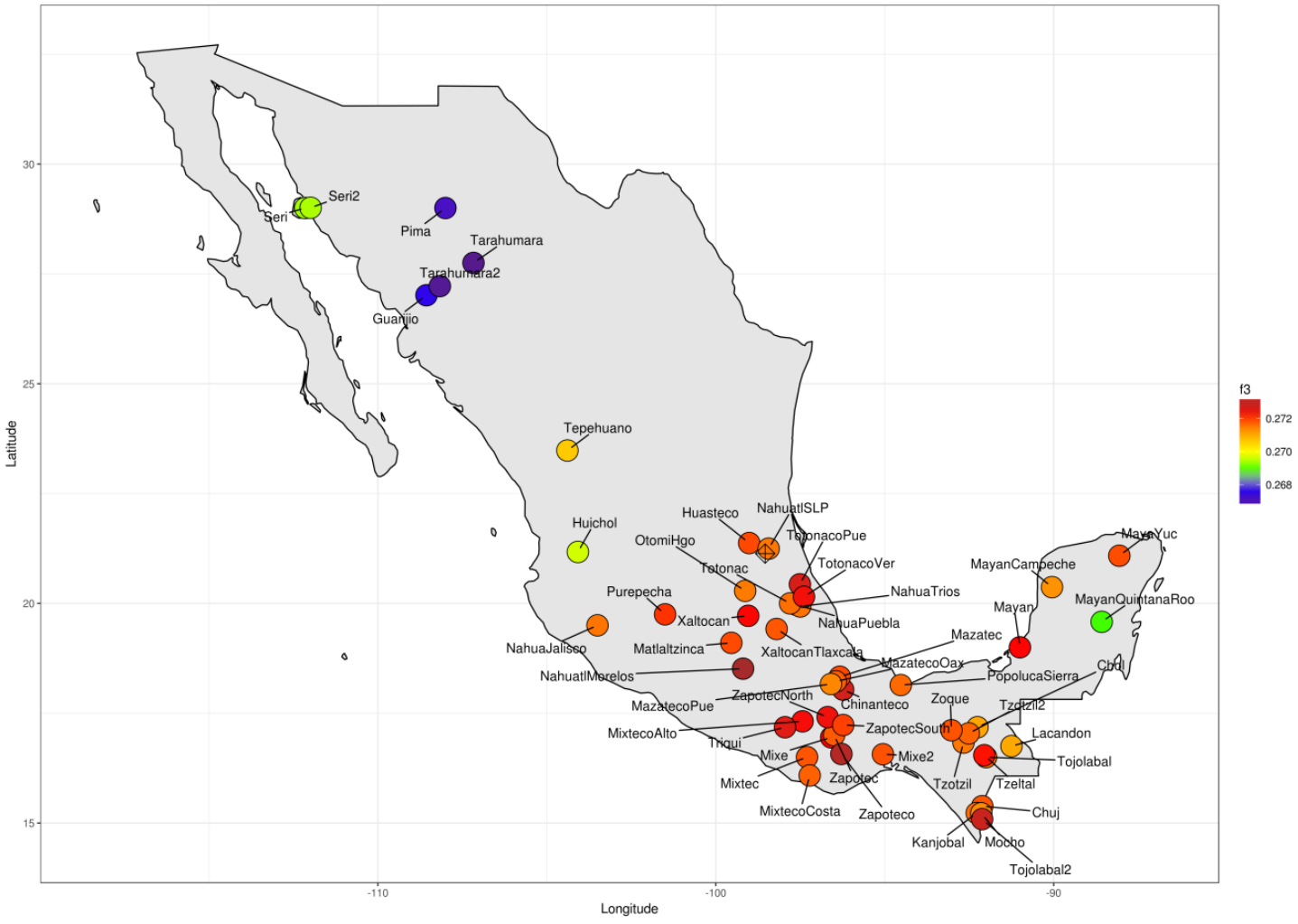
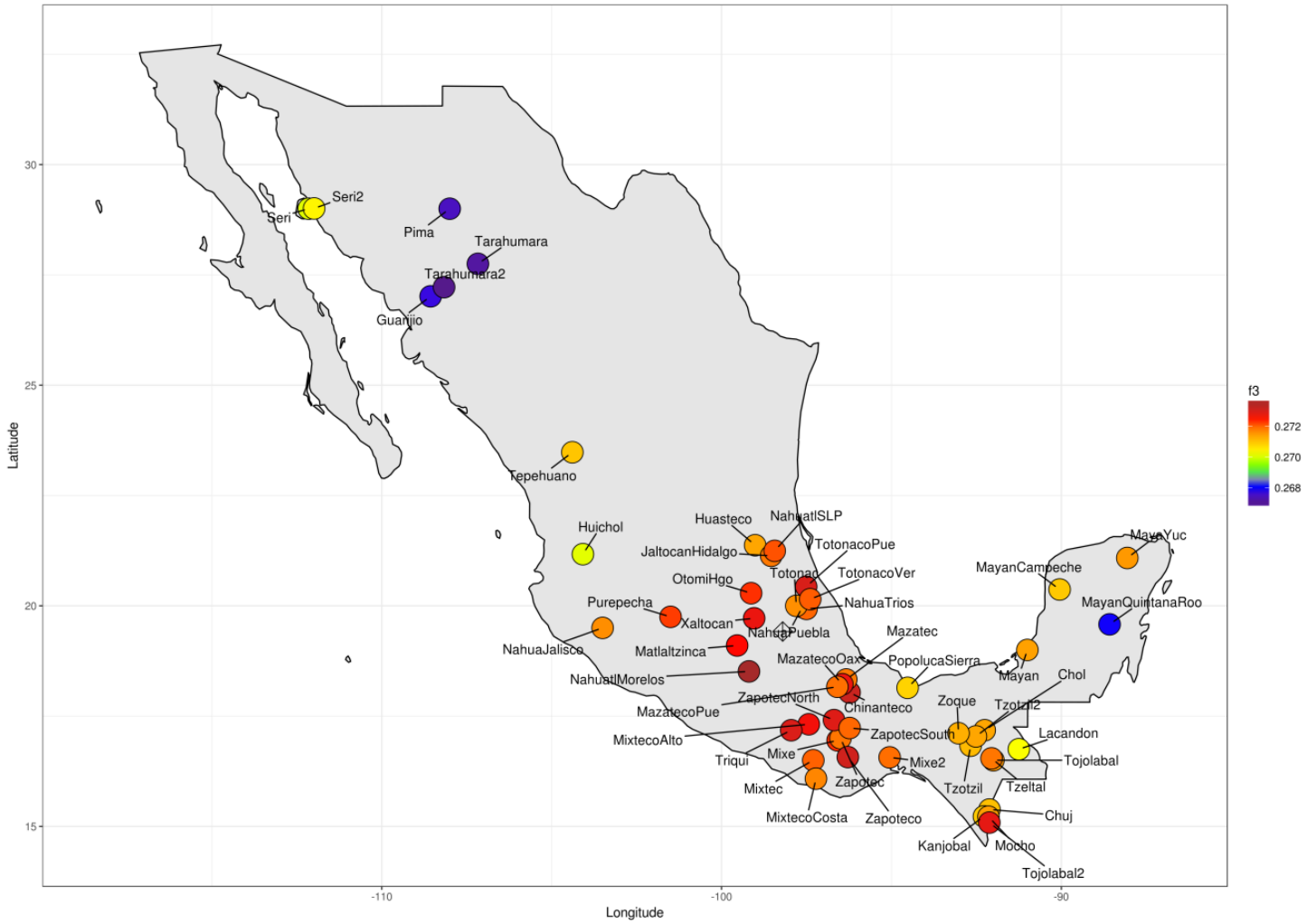


Figure 2.11 – Map of outgroup-f3 results for XaltocanTlaxcala. Warmer colors show a closer relationship to the population of interest.



the null hypothesis is zero, while positive values suggest that the PopulationX is closer to PopulationZ, and negative values suggest that the PopulationX is closer to PopulationY. Values with an absolute z-score higher than 3 were considered significant. The f4 statistics support a close relationship between the Central and South Mexican populations, with the only significant results in that region showing that Xaltocan is closer to NahuatlMorelos than a handful of other groups (**Table 2.7**). The majority of the significant f4 results, however, just show that the Xaltocans are closer to Central and South Mexican populations than they are to North or Southeast Mexican populations (**Table B2**).

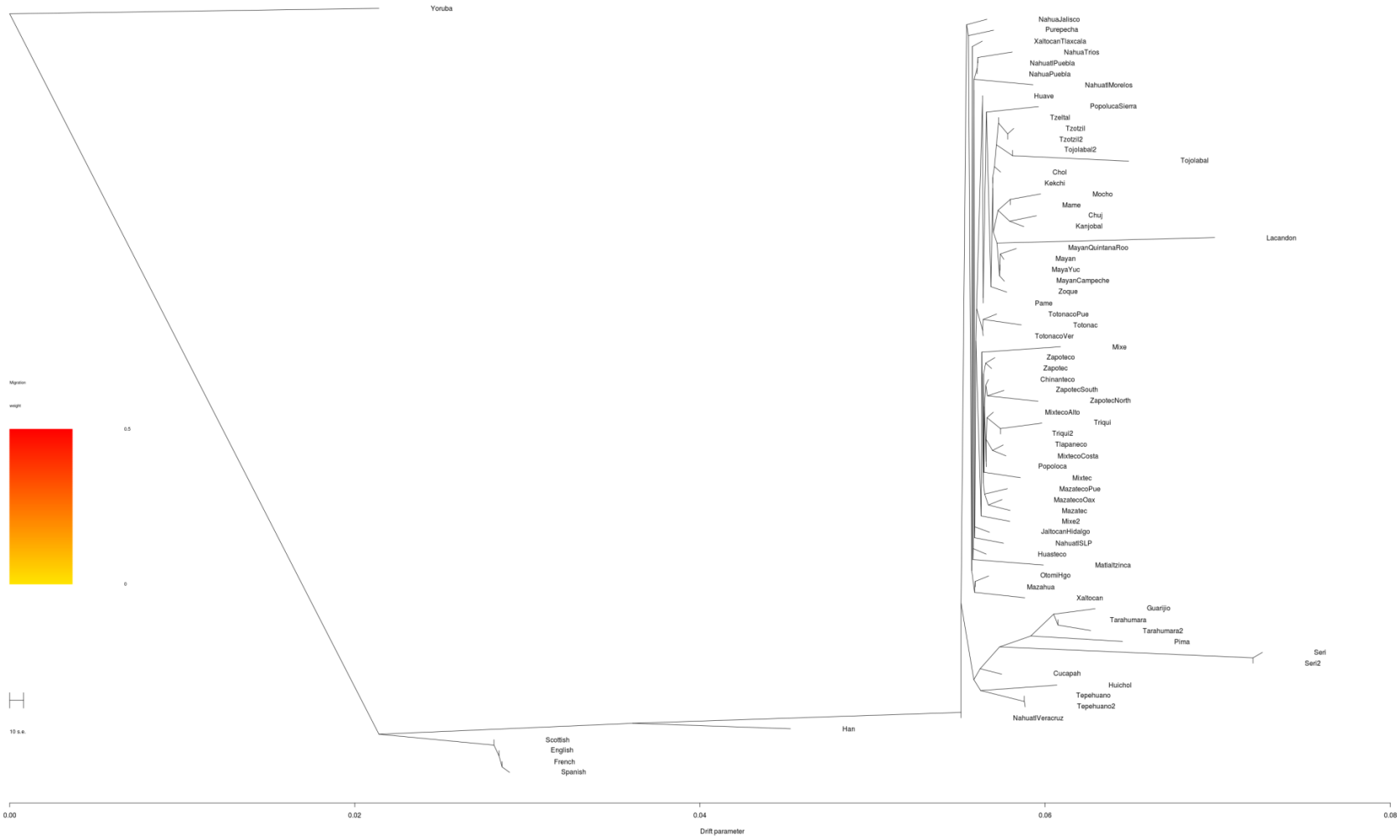
A maximum-likelihood tree constructed from the masked genotype data using Treemix (Pickrell and Pritchard 2012) with no admixture edges is in broad agreement with the patterns seen in the previous analyses (**Figure 2.12**). The tree shows the North Mexican groups forming their own clade, to the exclusion of the other populations. The Southeast populations form a clade with respect to the Central and South Mexican populations. The Xaltocan populations form clades with the populations most geographically proximate to them, not with each other.

Finally, I used the EEMS method (Petkova et al. 2016) to further explore how indigenous genetic diversity is geographically structured within Mexico. EEMS uses both genetic and physical distance to model effective migration rates across space, allowing for the recognition of general patterns of gene flow between populations. This tool has been used elsewhere to discover the geographic features impeding or

Table 2.7 – Significant f4 results within Central and Southern Mexico

YRI	X	Y	Z	f4	Zscore
Yoruba	Xaltocan	Huasteco	NahuatlMorelos	0.000638	3.053
Yoruba	Xaltocan	JaltocanHidalgo	NahuatlMorelos	0.000624	3.1
Yoruba	Xaltocan	Mazatec	NahuatlMorelos	0.00062	3.059
Yoruba	Xaltocan	NahuaJalisco	NahuatlMorelos	0.00066	3.14
Yoruba	Xaltocan	NahuaPuebla	NahuatlMorelos	0.000633	3.141
Yoruba	Xaltocan	Totonac	NahuatlMorelos	0.000638	3.092
Yoruba	Xaltocan	Zapotec	NahuatlMorelos	0.000693	3.147
Yoruba	Xaltocan	ZapotecSouth	NahuatlMorelos	0.000619	3.179
Yoruba	XaltocanTlaxcala	Huasteco	NahuatlMorelos	0.000712	3.181

Figure 2.12 – Maximum-likelihood tree of Mexican populations

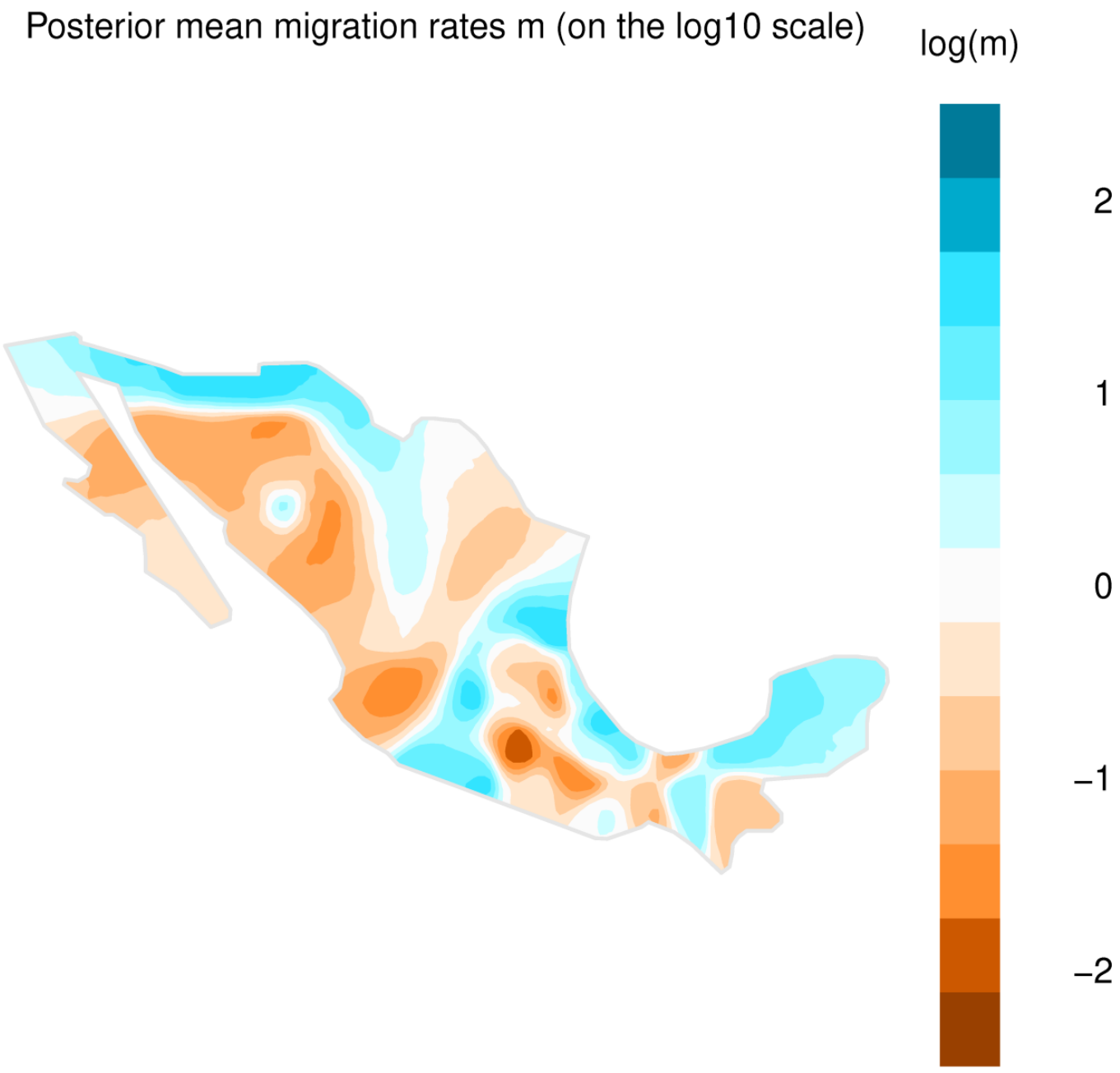


enabling gene flow within a region (Petkova et al. 2016; Peters et al. 2018). Here, it provides a more detailed look at the geographic distribution of genetic variation across Mexico and graphically illustrates the relationships between these groups. Areas of lower effective migration (**Figure 2.13**, shown in orange) in the EEMS analysis roughly correspond to the highland areas of Central and South Mexico, as well as the northern deserts in Sonora and Chihuahua.

To further explore the relationship between the three Xaltocan populations, I examined genomic regions shared identically by descent (IBD) between individuals within and between populations across Mexico. This comparison could only be done with individuals genotyped on the same array, however, because the overlap in SNPs genotyped across arrays was too low to allow for accurate IBD estimation. Because I was most interested in the possibilities of IBD sharing between the Xaltocans, I estimated the amount of shared IBD for the Mexican populations available on the Human Origins array (Pima, Mayan, Zapotec, Mixe, Mixtec, and the Xaltocans). I then split the segments into 9 categories corresponding to the length of segments shared and the degree of relation between individuals sharing them (Moreno-Estrada et al. 2014). Individuals from the same population tend to share more IBD tracts with other members of that population relative to individuals in another population, unless there has been gene flow between those populations.

I examined each bin of IBD tracts to see if the Xaltocans shared more IBD tracts with one another than with the other populations in the dataset, which would

Figure 2.13 – EEMS plot of effective migration rates across Mexico. Blue represents areas of higher effective migration while orange represents areas of lower effective migration.



suggest a closer relationship among these communities that could be due to the history of Xaltocan. Specifically, if all three Xaltocans share more IBD tracts with each other than with the other Mexican populations, it would be consistent with the hypothesis that some people fled Xaltocan after the Tepanec conquest and resettled in XaltocanTlaxcala and JaltocanHidalgo, while some others remained in the town, leaving descendants today. If Xaltocan shares more IBD tracts with either XaltocanTlaxcala or JaltocanHidalgo than with other Mexican populations, it may be indicative of some people abandoning the town after the Tepanec conquest, fleeing to XaltocanTlaxcala or JaltocanHidalgo but not both, with others remaining in the original town and leaving descendants there today. Lastly, if XaltocanTlaxcala and JaltocanHidalgo shared more IBD tracts compared to the other Mexican populations (including Xaltocan), it would be consistent with most (perhaps all) people abandoning the town after the Tepanec conquest, fleeing to XaltocanTlaxcala and JaltocanHidalgo. However, comparing the amount of shared IBD tracts within each of the nine bins, there is not a noticeable enrichment of shared IBD tracts between any of the three Xaltocans relative to other Mexican populations.

Modelling the Genetic Bottleneck After European Colonization

Taking into account the colonial history of Mexico and historical estimates that indigenous Mexican population sizes may have contracted up to 95% following European colonization (Ubelaker 2006), I estimated the parameters of this bottleneck with the data from unadmixed residents of Xaltocan using *fastsimcoal2*

(Excoffier et al. 2013). This method optimizes the composite likelihood of the observed joint site frequency spectrum (SFS) to estimate parameters of a given demographic model. As others have noted, demographic estimates from SFS calculated using genotype array data can be severely biased by the ascertainment scheme of the array (Clark et al. 2005). The Human Origins array, which was used to genotype the Xaltocan samples, was ascertained using several world-wide populations, making correcting for ascertainment bias relatively simple (Patterson et al. 2012). Following the example of Lipson et al. (2013), I limited SFS estimation to SNPs ascertained on the San panel, which should be evolving neutrally in populations outside of Africa. I first calculated the joint SFS between unadmixed individuals from Xaltocan and 30 Han Chinese individuals from the 1000 Genomes phase 3 dataset. I modified the model of Lindo et al. (2016) slightly to exclude accounting for European admixture in the Xaltocan population (**Figure 2.14**), given the lack of European ancestry present in this sample set, and then estimated both the timing and severity of the bottleneck. While previous analyses have had difficulty estimating the parameters of the bottleneck after European colonization in indigenous Mexican groups (Moreno-Estrada et al. 2014), the relatively larger sample size of the Xaltocan dataset may have more power to tease out these variables.

Table 2.8 shows the results of the estimated parameters of the highest likelihood model, as well as the confidence intervals. The estimated pre-bottleneck

Figure 2.14 – Visual representation of the demographic model. N_{AF} is the ancestral African population size (taken from Lindo et al. 2016). Dates of the Out-of-Africa bottleneck as well as the bottleneck corresponding to the peopling of the Americas were also taken from Lindo et al. 2016. CHB is the Han Chinese population from 1000 Genomes Project (2013).

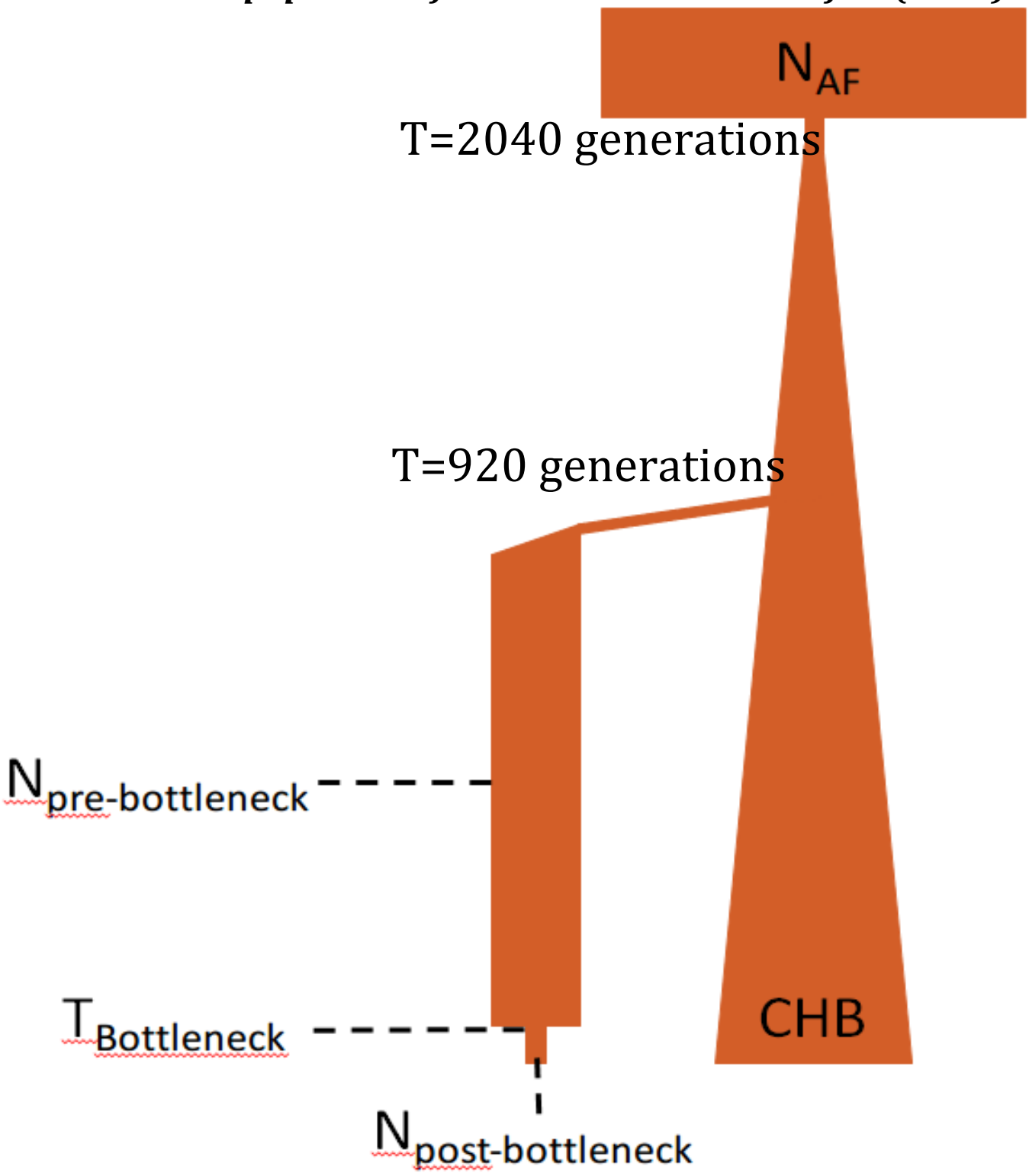


Table 2.8 – fastsimcoal2 bottleneck parameter estimates with (confidence intervals) of the best likelihood model.

Population size (Ne) post-bottleneck	Population size (Ne) pre-bottleneck	Time of bottleneck (generations)
6249 (2427-16886)	8909 (7782-10534)	6 (4-20)

effective population size from the best-likelihood model is larger than post-bottleneck effective population size, indicative of a bottleneck in the Xaltocan populations. However, there is substantial overlap between the confidence intervals of the estimated pre- and post-bottleneck population sizes at Xaltocan, suggesting that the effective population size for many of the simulated datasets actually remained the same or increased after the bottleneck events. This means that either there was not a recent bottleneck at Xaltocan or there is not enough power using the current data to detect the bottleneck.

Discussion

The genetic impacts of European colonialism vary widely across the Americas, but have generally been treated not as a subject of interest, but rather as a complication inhibiting our understanding of the population history of the region before European colonization, or as an obstacle in studies seeking to understand the genetic background of some medically relevant trait. In contast, in this study of Mexican populations, I leverage both the indigenous genetic components and the genetic components derived from recent European and African admixture events to better understand the history of Mexico. Using old and new genome-wide datasets from across Mexico, as well as the 96 Xaltocan samples, I was able to make several novel insights about the history of admixture in Mexico and the geographic structure of genetic variation in indigenous Mexican groups, as well as test hypotheses about the relationships among the three Xaltocans.

Notably, admixture estimates for the indigenous populations of Mexico are quite low compared to other datasets in the Americas. The highest levels of European admixture occurred in Central Mexico, and the earliest admixture events appear to have occurred in South Mexico and Central Mexico. These results are consistent with historical evidence, as they match the areas where the Spanish first arrived and concentrated their power within colonial Mexico. The relatively higher amounts of African ancestry in Southeast Mexico is consistent with recent historical work suggesting that there were many more African slaves in the region than previously realized (Restall 2009). Estimates for the timing of admixture in Xaltocan (late 18th-early 19th centuries) are also consistent with the historical record for this town. Early colonial documents suggest that the town was run by the indigenous inhabitants and rarely visited by Spanish colonists or enslaved Africans during the early colonial period (Hicks 2005). Estimates for XaltocanTlaxcala and JaltocanHidalgo indicate similar dates of admixture, suggesting that it may have taken longer for Spanish colonists to spread from urban areas of Mexico into the countryside.

Variation within the Native American genetic components of indigenous Mexican groups broadly reflects the geographic differences between these groups. The PCA and ADMIXTURE plots show strong northwest to southeast trends, which have been seen before in this region (Moreno-Estrada et al. 2014). Furthermore, in the model of spatial genetic variation provided by EEMS, it appears that the

highlands of Mexico, as well as the Sonoran and Chihuahuam Deserts in the north of the country, provide significant barriers to gene flow between human populations. Overall these results support the hypothesis that the geographic landscape has played a significant role in shaping indigenous genetic diversity in Mexico.

No Close Genetic Relationship Between the Xaltocans

Oral histories suggested a possible genetic link between the three Xaltocans. However my analysis of genetic data from the towns has not revealed a close relationship between them relative to other populations in the region. The *f-statistics* show a close relationship between all of the populations in Central and South Mexico, but the Xaltocans are not significantly closer to each other. The maximum-likelihood tree also suggests that the Xaltocans are closer genetically to the populations that are closest to them geographically, rather than closest to each other. These results are perhaps not surprising if the ancient residents of Xaltocan actually did abandon the town after the Tepanec conquest, only to be replaced during the Aztec period. In that case, I would not expect to see a close relationship between the modern residents of Xaltocan and those of XaltocanTlaxcala and JaltocanHidalgo. However my analysis shows close relationships among most modern groups in Central and South Mexico, suggesting there may have been so much gene flow between groups in this region that it masks any particularly close relationship between the Xaltocans.

Challenges in Estimating Bottleneck Parameters

Despite historic evidence of severe demographic crashes in indigenous populations in Mexico, this study provides no clear evidence of a recent bottleneck in the Xaltocan population. There are several possible explanations for this. First, because this dataset was genotyped on a the Human Origins array, I had to limit the SNPs used to calculate the SFS to those in a single population panel to control for the effect of ascertainment bias on the distribution of the SFS before estimating the parameters of the bottleneck. The genotyping array is likely also missing the full range of variation present in the population because, while the Human Origins array was designed for use in world-wide populations, genetic diversity in the Americas is not well represented on the array (the only American population used in the SNP selection, the Karitiana, contributed only ~2000 SNPs to the final array). Both of these factors limit my power to detect demographic changes at Xaltocan.

Second, it is also possible that the bottleneck happened so recently that coalescent-based methods such as fastsimcoal2 will have problems detecting it. Because fastsimcoal2 is a coalescent modeling program and the likely bottleneck occurred in the past ~20 generations, there may not be enough coalescent events with 140K SNPs to detect a bottleneck in this timescale. Another study (Lindo et al. 2016) was able to detect a recent bottleneck in a similar demographic scenario in the Americas, with similar sample size but many more SNPs (~2.7Mb), suggesting

that an increase in SNPs in the Xaltocan dataset may benefit future demographic modeling efforts.

Lastly, it is possible that the Xaltocan community today may contain individuals drawn from multiple different pre-contact communities. The Spanish in particular were known for gathering different indigenous groups together into *congregaciones* in an attempt to increase tribute, labor, Spanish rule of law, and Christianity in these groups (Gibson 1964; Lockhart 1992; Endfield 2008). Xaltocan may have experienced substantial gene flow during the colonial period associated with this movement of people around Central Mexico, increasing genetic diversity at this location and masking any signal of a bottleneck that may have occurred. The close relationship between modern groups in Central and South Mexico does not currently allow us to pull apart these potentially different ancestral components, but whole genome sequencing and further work on ancient DNA from the region may help us better differentiate among populations and tell us more about whether extensive genetic interactions occurred in this region before the Spanish conquest, or if this pattern of close genetic relations seen in the region today is the result of processes triggered by colonization.

Conclusions

This study highlights the importance of combining genomic studies with historic and archaeological data to obtain new insights into the population history of a region. Here, I was able to show significant differences in the admixture

proportions and the timing of admixture events across Mexico, reflecting the complicated colonial history of the country. Furthermore, I was able to show that geographic features such as deserts and highlands were in fact barriers to gene flow among the indigenous groups of Mexico. The current analysis was unable to detect a genetic bottleneck in the Xaltocan population, but future work with an expanded dataset may prove more successful. Overall, this study greatly extends our understanding of colonial interactions in the Americas and provides a blueprint for future studies of regional genetic variation.

Methods

Samples, DNA extraction, and Genotyping

Saliva samples were collected from 93 residents of the towns of Xaltocan, State of Mexico (N=47), Xaltocan, Tlaxcala (N=24), and Jaltocan, Hidalgo (N=22). Community authorities and representative bodies were consulted where appropriate, and all individual participants provided written informed consent for the types of analyses conducted in this study. The collection and analysis of samples from all communities was approved by the Institutional Review Board of the University of Texas at Austin (protocol #2012-05-0105). I extracted DNA from the samples using the DNA Genotek prepIT-L2P protocol. Samples were genotyped for 627,151 SNPs of the Human Origins Array at the Children's Hospital of Philadelphia's Center for Applied Genomics. The raw genotypes were QC'd and

combined with other datasets genotyped on this platform by David Reich at Harvard University.

Merging with Modern Comparative Datasets

I combined the Human Origins array dataset with three comparative datasets of individuals from around Mexico, including cosmopolitan cohorts and indigenous groups, for further analysis (Moreno-Estrada et al. 2014; Orozco et al. in prep). Individuals with >10% missing data and SNPs with >5% missingness were removed from the dataset using PLINK v1.9 (Chang et al. 2015). I used the program *smartrel* from the EIGENSOFT package (Patterson et al. 2006) to identify and remove individuals from the dataset who were second-degree relatives or closer. The remaining individuals were phased using SHAPEIT2 with default parameters and the 1000 Genomes Phase 3 dataset as a reference panel (Delaneau et al. 2012; 1000 Genomes Project Consortium 2015). I next used RFMIX (Maples et al. 2013) to assign each chromosomal segment to its most likely ancestral source for each Native American individual in the modern dataset. I used 30 YRI individuals, 30 CEU individuals, and 30 unadmixed PEL individuals from the 1000 Genomes Phase 3 dataset (1000 Genomes Project Consortium 2015) to represent the possible ancestral populations for this local ancestry assignment. After ancestry assignment, I masked SNPs from the data for an individual if a non-indigenous allele was identified at that locus or if that locus had a low-confidence (<90%) ancestry assignment as missing. I then removed any individuals with >25% missing SNPs,

except for the XaltocanTlaxcala population as this would have removed all of those individuals (a 50% missing data cutoff was used for this group). I then merged all of the arrays for a final dataset of 644 indigenous Mexican individuals and 102,176 SNPs for comparative analysis. Detailed information about the comparative dataset can be found in **Table B1**.

ADMIXTURE, Treemix, and Principal Component Analysis

I performed model-based clustering analysis using ADMIXTURE (Alexander et al. 2009) on the combined dataset before masking. First I pruned the dataset by linkage disequilibrium using PLINK (Chang et al. 2015) with the flag –indep-pairwise 200 1 0.4, leaving 71832 SNPs. I ran admixture with the cross validation flag from K=2 to K=15 clusters (**Figures 2.15-16**), with 10 replicates for each value of K. For each K value, the replicate with the highest log likelihood was kept. I performed principal component analysis of the pruned datasets using the *smartpca* program in EIGENSOFT (Patterson et al. 2006). I constructed a maximum-likelihood tree of the masked indigenous Mexican populations using Treemix v1.13 (Pickrell and Pritchard 2012).

***f*-statistics**

I computed *f*-statistics on the combined dataset using ADMIXTOOLS (Patterson et al. 2012). I used *qpDstat* with *f4mode* for *f*4-statistics and *qp3Pop* for outgroup *f*3-statistics. I computed standard errors using a weighted block jackknife over 5-Mb blocks.

Figure 2.15 – Cross-validation errors for ADMIXTURE runs K=2-15.

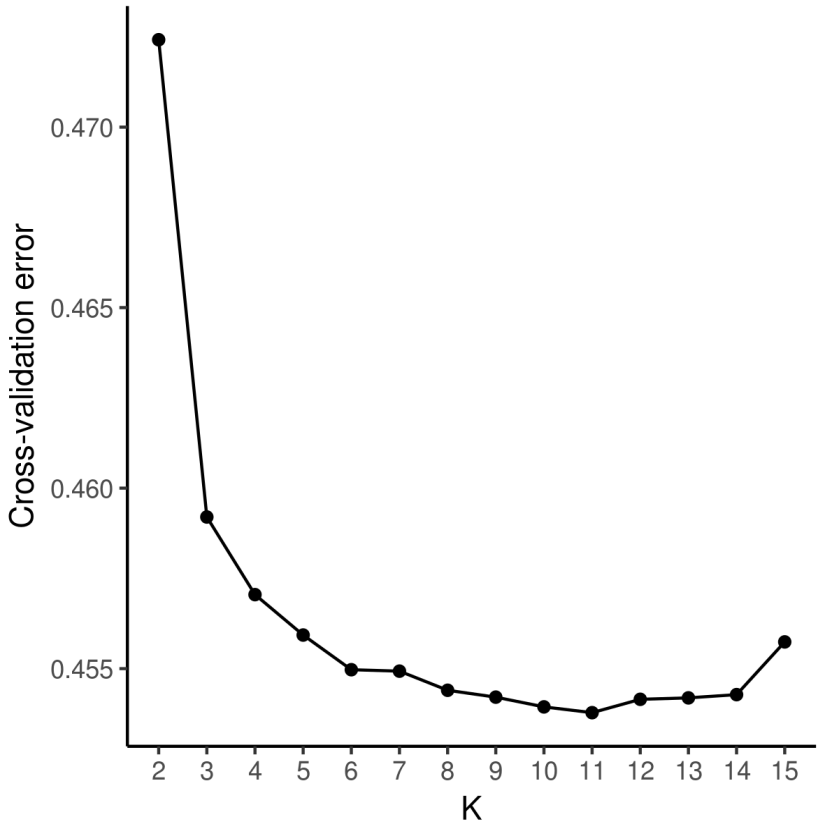
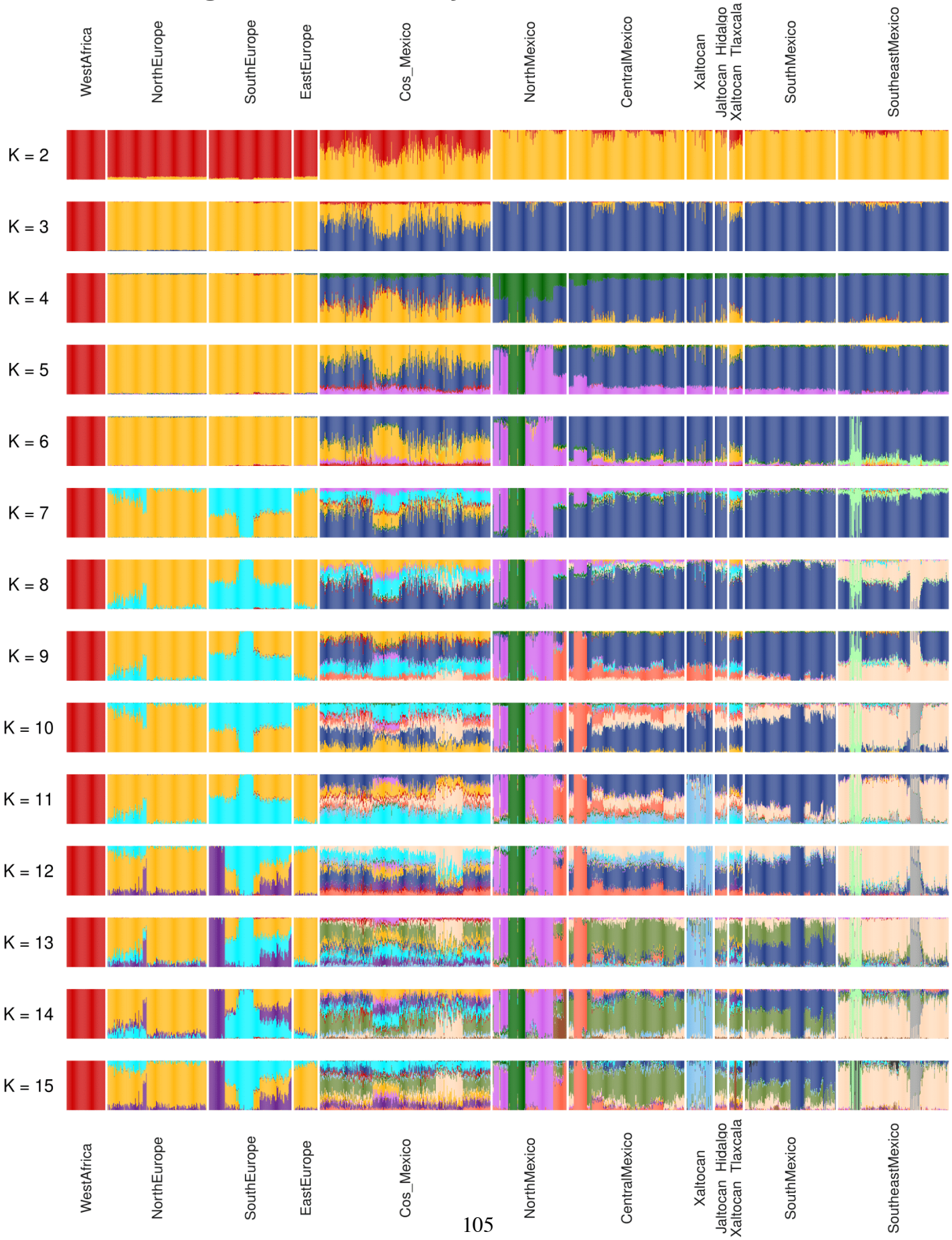


Figure 2.16 – Plot of all ADMIXTURE runs.



Identity-by-descent (IBD) analysis

I estimated the amount of DNA shared identically by descent (IBD) using the GERMLINE software (Gusev et al. 2009), with a 5-cM threshold to minimize false positive IBD matches. All 5 cM or greater segments of shared IBD between pairs of individuals were summed, and binned into nine categories corresponding to degrees of relatedness following Moreno-Estrada et al. (2014).

Alder analysis

I used the ALDER program to estimate the timing of admixture events in the three Xaltocan populations using 30 CEU, YRI, and unadmixed PEL individuals respectively from the 1000 Genomes Phase 3 dataset as potential source populations. Alder uses weighted LD curves between a potentially admixed population and user-defined reference populations to estimate the time and extent of an admixture event (Loh et al. 2013).

Analysis of spatial genetic structure in Mexico

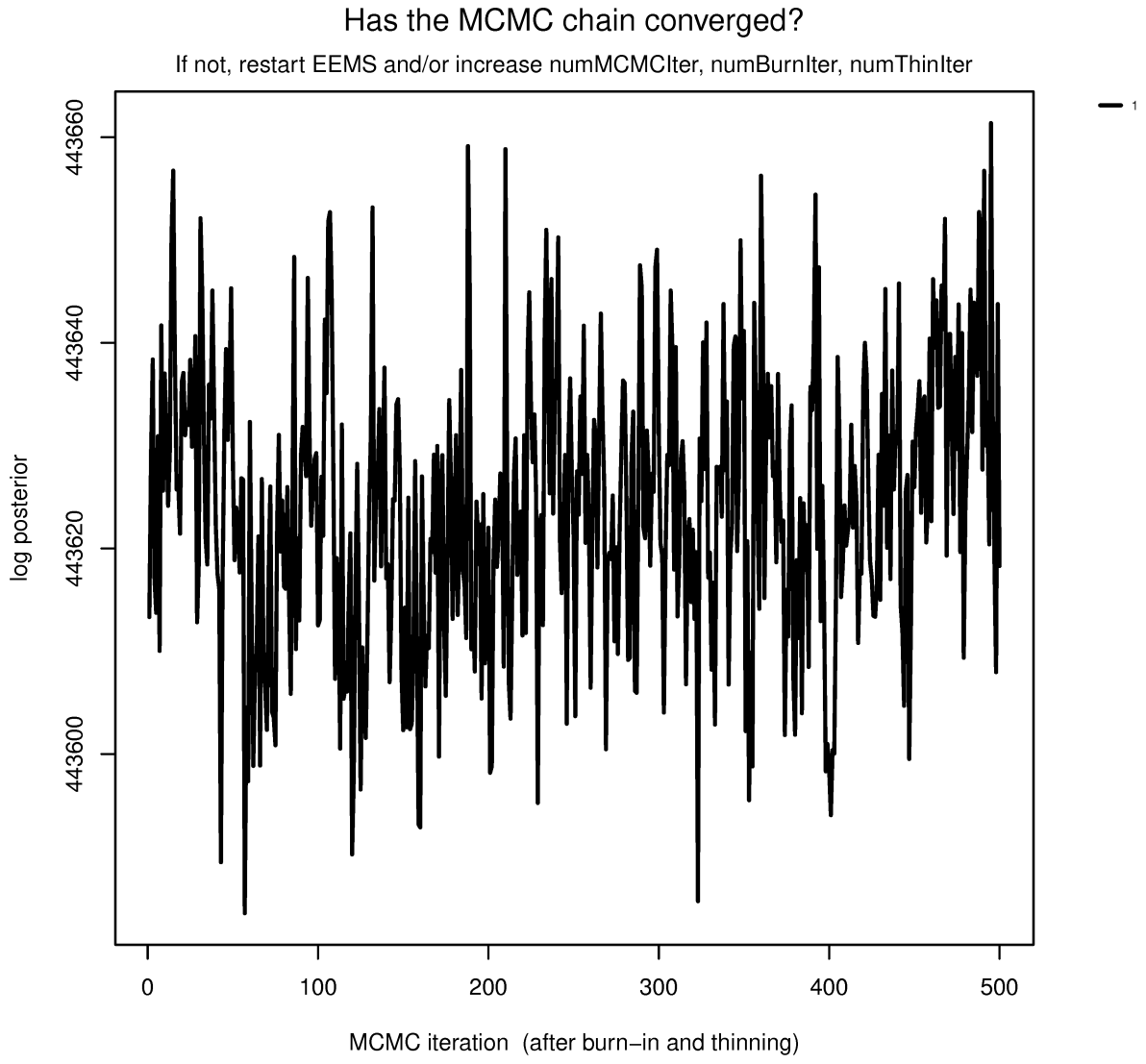
To better understand the genetic variation across Mexico and to identify any potential geographic barriers to gene flow between populations, I used the EEMS method (Petkova et al. 2016). EEMS uses both genetic and physical distance to model "effective migration" rates across space, allowing for the recognition of general patterns of gene flow between populations. An effective migration surface then represents migration rates that, within an idealized stepping stone model evolving under equilibrium, would produce genetic dissimilarities similar to those

observed in the data. Because EEMS is sensitive to missing data, I further pruned my masked dataset of indigenous Mexican individuals with <8% missing data and SNPs with <4% missingness, leaving 89,939 SNPs and 491 individuals across Mexico for analysis. Following the example of Peter et al (2018), I ran several test runs with different parameters to get the acceptance proportions of my parameters into acceptable ranges according to the documentation for the program. The only parameter that needed to be altered from the default settings was *mrateMu*, which was set to 0.001. I ran six pilot runs with 250 demes for 6 million iterations and a 1 million iteration burn-in. I chose the run with the best likelihood and checked its tracer to make sure that the parameter space had been sufficiently explored before carrying out further analysis (**Figure 2.17**).

Bottleneck parameter estimates using fastsimcoal2

To estimate the parameters of the potential bottleneck at Xaltocan, I constructed a demographic model using fastsimcoal2 (Excoffier et al. 2013). Because the Human Origins Array data is ascertained by choosing SNPs variable in the two chromosomes of an individual of known ancestry from one of 13 population panels, one can avoid the bias that this might introduce into demographic estimates by limiting your analysis to the SNPs from a single panel. I therefore limited the SNPs in my dataset to the 140,000 in the San panel as this allowed the largest number of SNPs to be used in the analysis. I used Arlequin to calculate the site frequency spectrum (SFS) of my dataset (Excoffier and Lischer 2010). Following the

Figure 2.17 – Tracer plot for the EEMS run, showing that parameter space has been well explored.



Chapter 3: Comparing Signals of Natural Selection Between Three Indigenous North American Populations

Abstract

While many studies have highlighted human adaptations to diverse environments worldwide, genomic studies of natural selection in Indigenous populations in the Americas have been absent from this literature until very recently. Since humans first entered the Americas some 20,000 years ago, they have settled in many new environments across the continent. This diversity of environments has placed variable selective pressures on the populations living in each region, but the effects of these pressures have not been extensively studied to date. To help fill this gap, I collected genome-wide data from three Indigenous North American populations from different geographic regions of the continent (Alaska, Southeastern United States, and Central Mexico). I identified signals of natural selection in each population and compared signals across populations to explore the differences in selective pressures among the three populations sampled. I find evidence of adaptation to cold and high-latitude environments in Alaska, while in the Southeastern US and Central Mexico, pathogenic environments appear to have created important selective pressures. This study lays the foundation for further functional and phenotypic work on possible adaptations to varied environments during the history of population diversification in the Americas.

Significance Statement

Recent studies have shown that humans have adapted to many different environments around the world. However, few of these studies have centered on Indigenous groups in the Americas. I present the first comparative analysis of genetic adaptations in populations across North America using genome-wide scans for signals of natural selection in three populations from Alaska, the Southeastern United States, and Central Mexico, each inhabiting vastly different environments. I find evidence for adaptation to cold and high latitudes in the Alaskan population; whereas infectious disease was a strong selective pressure in the Southeastern US and Central Mexico. Because there are few shared signals of selection between populations, these sweeps likely occurred in the last 10-15 thousand years (ky), after population differentiation in the Americas. This study fills an important gap in my knowledge of genetic adaptations in humans to various environments.

Introduction

Since first leaving their ancestral environments in Africa more than 100,000 years ago, humans have spread to nearly every region of the planet. In doing so, different populations have been exposed to many new environments and selective pressures, and have developed a diversity of adaptations as a result (Fan et al. 2016). The declining cost of array and sequencing technologies and the improvement of methods for detecting signals of natural selection have allowed

researchers to answer questions about selective pressures across a growing number of populations worldwide (Perry et al. 2014; Karlsson et al. 2013; Yi et al. 2010).

However, very little is known about the recent evolutionary history and selective pressures encountered by Indigenous North American populations. Indigenous North American populations are underrepresented in the population genetics literature as a whole (Bolnick et al. 2016) and in studies of selection in particular. Only a handful of genomic studies of natural selection have been conducted in the Americas, and the majority of these have focused on populations in South America (Pickrell et al. 2009; Bigham et al. 2010; Schlebusch et al. 2015; Mychaleckyj et al. 2017). To my knowledge, only two genomic studies of selection have been published in North American populations. Lindo et al. (2016) found evidence of a complex history of selective pressures on the immune gene *HLA-DQ1* using exome data from ancient and modern populations in the Canadian Pacific Northwest. Another study with the Greenlandic Inuit found evidence of selection in the *FADS* genes, which code for fatty acid desaturases that are associated with polyunsaturated fatty acid levels in the blood (Fumagalli et al. 2015), as well as in the genomic region encompassing *TBX15,* which plays a role in the differentiation of brown and white adipocytes. The authors suggest these signals of selection are likely related to adaptation to cold environments.

Here I present the first genome-wide scans for natural selection across three populations from different regions of North America. I find evidence of adaptation to

cold and high latitudes in an Alaskan Native population from the Arctic, and evidence of selection at several genes related to inflammation and immune function in populations from the Southeastern United States and Central Mexico. I find little overlap between putatively selected genes in these three populations, suggesting that local selective pressures in each geographic region have shaped these indigenous North American populations differently since they diverged after settling in distinct regions of the continent.

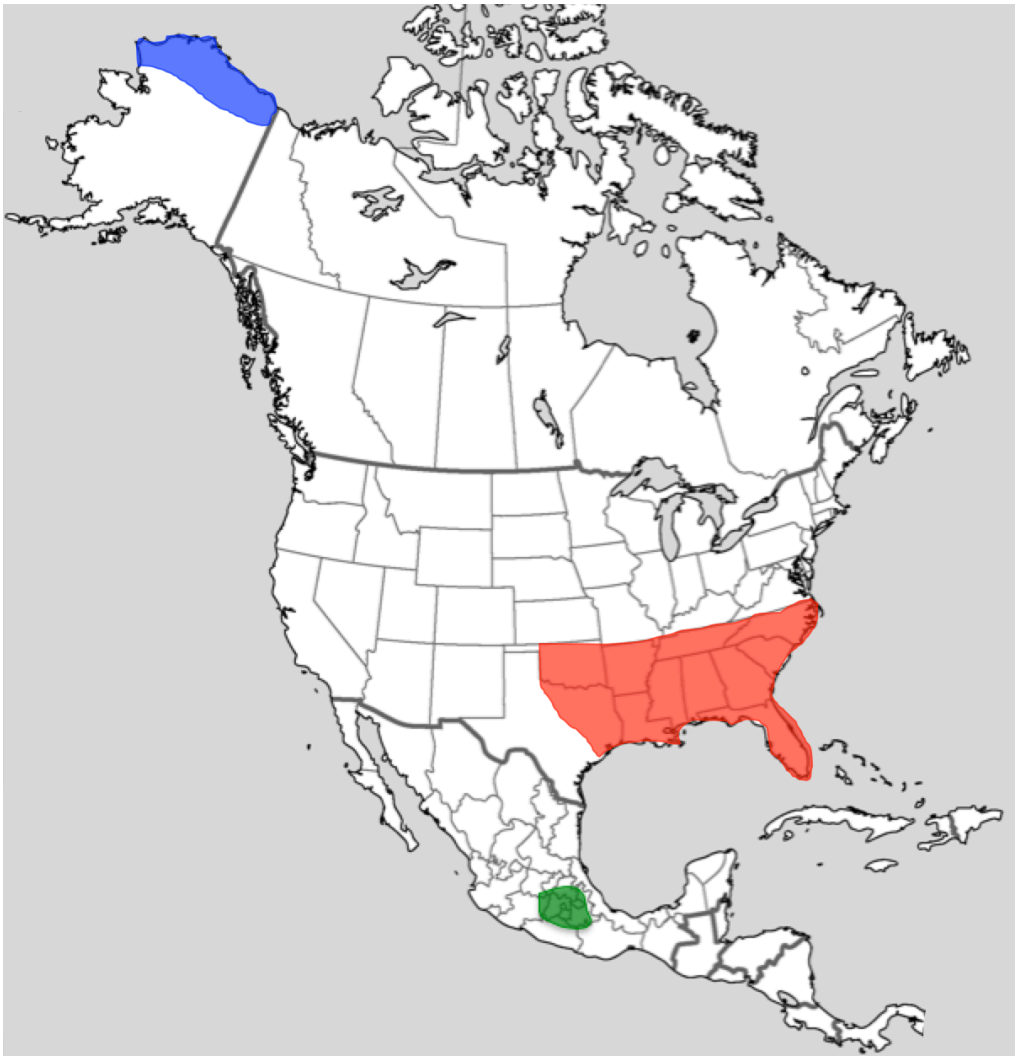
Results

Data Collection and Genetic Ancestry Estimates

I collected DNA samples from 150 individuals from three Indigenous North American populations (**Figure 3.1**), including 35 Alaskan Iñupiat from the North Slope of Alaska, 47 individuals from the town of Xaltocan in Central Mexico, and 68 individuals from several closely related communities in the Southeastern United States (populations referred to hereafter as Alaska, Central Mexico, and Southeastern US, respectively). I then used the Affymetrix Axiom Human Origins Array to genotype 629,443 genome-wide single nucleotide polymorphisms (SNPs) for each of these individuals. A total of 563,162 SNPs were included in my analyses after quality control filtering and merging with the 1000 Genomes dataset (1000 Genomes Project Consortium 2015) for comparative analyses.

Because many previously genotyped Indigenous populations of the Americas trace a large percentage of genetic ancestry to recent European and African

Figure 3.1: Map of sampling areas. Blue = Alaska, Red = Southeastern US, Green = Central Mexico. Specific sampling locations, where publicly available, are provided in Supplemental Figures 1 and 2.

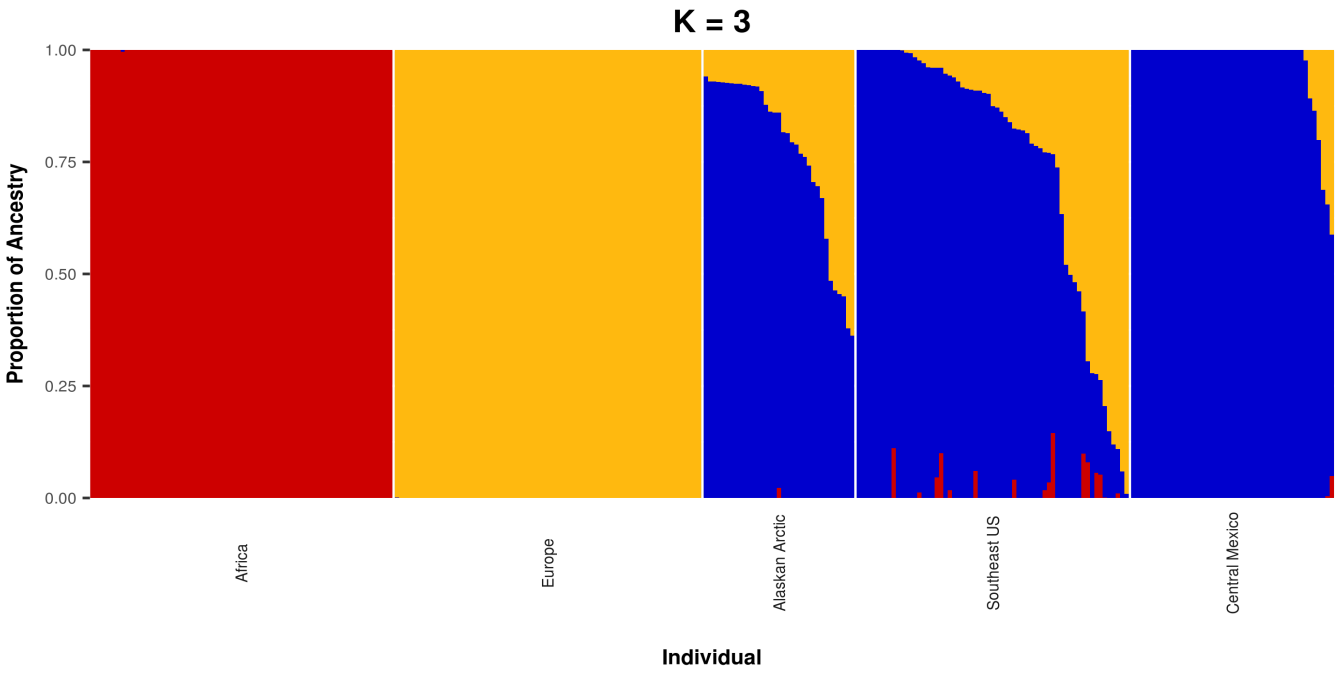


ancestors, which can influence results of genome-wide scans for selection, I first conducted a nonhierarchical clustering analysis of the SNP data implemented in the program ADMIXTURE (Alexander et al. 2009). **Figure 3.2** shows that many of the Alaska and Southeastern US individuals have more European ancestors than the individuals sampled in Central Mexico. Using the ADMIXTURE K=3 results, I restricted further analysis to individuals with more than 80% cluster membership in the American component to maximize my power to detect signals of selection while minimizing the effects of admixture on my results.

Genome-wide Scans for Signals of Natural Selection

I computed two statistics to identify potential signatures of natural selection in the three study populations after restricting my datasets to individuals with <20% non-indigenous ancestry. First, I calculated the Population Branch Statistic (PBS) for each autosomal SNP in each population using individuals from the 1000 Genomes Peruvian population without recent European or African ancestry as an ingroup and the 1000 Genomes Han Chinese population as an outgroup. PBS computes the amount of genetic differentiation at a given locus along a branch leading to a population of interest by comparing transformed pairwise-F_{ST} values between each pair of three populations (Yi et al. 2010). A population's PBS value at a given locus corresponds to the magnitude of allele frequency change relative to its divergence from the other two populations. This approach has proven to be powerful at detecting recent signals of selection (Yi et al. 2010; Lindo et al. 2016). I

Figure 3.2: *ADMIXTURE* analysis of population structure.



also calculated the integrated haplotype score (iHS), a widely used haplotype-based method of detecting signals of selection, for each autosomal SNP in each of the three study populations. The iHS is a measure of extended homozygosity in the haplotype surrounding a given SNP. Extended stretches of homozygosity, relative to the background is a signal of a selective sweep that has not yet reached fixation. I then identified the top 1% of PBS and iHS scores from each population and cross-referenced them to find SNPs that were in the top 1% of both statistics. This approach should reduce my chances of reporting false positives, as iHS has been shown to be robust to demographic history that is often a confounding factor in F_{ST}-based approaches such as PBS (Voight et al. 2006).

I found 66 putatively selected SNPs for the Alaska population, 63 such SNPs for the Southeastern US population, and 112 such SNPs for the Central Mexico population. A subset of genes with the strongest signals of selection for each population are listed in **Table 3.1**. A complete list of the putatively selected genes for each population is available in **Table C1**.

Functions of the Putatively Selected Genes

I used GeneCards (Stelzer et al. 2016) to gain insight into the functions of the putatively selected genes. The strongest signals of selection in the Alaska population occur within genes related to vitamin D metabolism (*CYP2R1*), heparan sulfate biosynthesis (*HS3ST4*), adipose tissue production (*KCNH1*), energy production (*VMA8*), and melanin production (*OCA2*). The genes with the strongest evidence of

Table 3.1 : Genes with the strongest signals of selection in each population. The listed PBS and |iHS| values are averaged over significant SNPs in each gene.

Population	Gene	PBS	|iHS|
Alaska	*CYP2R1*	0.704	2.418
	HS3ST4	0.451	2.883
	KCNH1	0.506	3.126
	OCA2	0.448	3.13
	VMA8	1	2.903
Central Mexico	*CNIH3*	0.411	2.587
	CNTN1	0.246	2.927
	IL1R1	0.238	3.147
	MUC19	0.254	2.803
	OTOF	0.339	2.951
Southeastern US	*A4GALT*	0.29	3.594
	CNTN1	0.22	3.442
	IL1R1	0.248	3.281
	SLIT2	0.418	3.277
	VMP1	0.308	3.74

selection in the Southeastern US are related to immune response (*A4GALT*, *IL1R1*), cell death (*VMP2*), and cellular migration and adhesion (*SLIT2, CNTN1*). The genes with the strongest evidence of selection in Central Mexico are related to immune response (*IL1R1, MUC19*), cell adhesion (*CNTN1*), ion transmembrane receptors (*CNIH3*), and hearing (*OTOF*).

I also conducted a pathway enrichment analysis using the WebGestalt platform (Wang et al. 2017) to better understand the metabolic pathways involving putatively selected genes and to determine if these genes were overrepresented in any particular pathways. After correcting for the false discovery rate, I found no significant enrichment of the putatively selected genes in any metabolic pathways.

However, I did see some interesting patterns in the metabolic pathways containing the genes of interest. In the Alaska population, I identified a number of putatively selected genes related to water reabsorption and carbohydrate metabolism. I also found that a number of putatively selected genes in both the Southeastern US and Central Mexico populations are related to inflammatory processes (**Tables C2-3**).

Shared Signals of Selection Between Populations

I next looked for shared signals of selection among the three study populations by comparing statistically significant results from each analysis. I found no shared signals of selection among all three populations, but did identify some putatively selected genes shared between pairs of populations. I found one shared

signal of selection at the gene *SLIT2* in the Alaska and Southeastern US populations. I also found that the Southeastern US and Central Mexico populations share signals of selection in seven genes (*MCHR1*, *CNTN1*, *IL1R1*, *LRRK2*, *TNRC6B*, *SLC9A2*). Among these genes, all but *CNTN1* are related to immune system pathways. I found no shared signals of selection between the Alaska and Central Mexico populations. The greater percentage of putatively selected genes shared between the Southeastern US and Central Mexico may be due either to similar selective pressures on both populations or to the more recent divergence of the Southeastern US and Central Mexico populations if the selective pressures primarily occurred before their split.

Discussion

Our results suggest that different selective pressures have been acting on the three study populations sampled from different regions of North America. First, in the Alaska population, I see evidence for adaptation to both cold and high-latitude environments. Two of the genes with the strongest signals of selection in this population (*CYP2R1, OCA2*) play a role in vitamin D metabolism. Vitamin D is an essential nutrient important for skeletal development and the innate immune response, among other processes. In humans, the majority of vitamin D synthesis takes place in the skin as a result of the interaction between cholesterol and UVB radiation from sunlight. High latitude regions, such as Alaska, are exposed to much lower levels of UVB radiation than other parts of the globe, making it difficult for people living in these areas to maintain healthy levels of vitamin D (Jablonski and

Chaplin 2010). The *CYP2R1* gene, which I find to be under selection in the Alaska population, is directly involved in the production of vitamin D (Cheng et al. 2004) and variants in the gene are associated with serum vitamin D levels (Wang et al. 2010; Ahn et al. 2010; Jiang et al. 2018). *OCA2*, also under selection in the Alaskan Arctic population, is associated with skin/eye/hair pigmentation and is indirectly involved in vitamin D metabolism as darker skin pigmentation protects against solar radiation, thereby mediating vitamin D production. Previous work has shown that variation in the *OCA2* gene is correlated with the amount of winter solar radiation (Hancock et al. 2011).

Other genes found to be under selection in the Alaska population may also be responses to Arctic environments. The gene *HS3ST4* is involved in the production of heparan sulfate, a molecule that affects blood thickness. Previous work has shown that extended exposure to cold temperatures increases blood thickness, increasing the risk of cardiovascular problems (De Lorenzo et al. 1999), so it is not surprising to see evidence of selection on this gene in this population. The gene *KCNH1* is involved in the regulation of cell proliferation and differentiation, in particular adipogenic and osteogenic differentiation in bone marrow-derived stem cells (Zhang et al. 2014). The *VMA8* gene, also known as *ATP6V1D*, produces a subunit of vacuolar ATPase, which produces much of the energy necessary for intracellular processes (Stevens and Forgac 1997). Thus, both of these putatively selected genes are involved in fat and energy production, suggesting that a cold-resistant

phenotype may be under selection in this Alaska population. Our pathway analysis also found a number of selected genes involved in water reabsorption and metabolism, which may have been advantageous in the Arctic environment as well. Previous studies of other Arctic populations have similarly found signals of selection related to metabolic pathways (Fumagalli et al. 2015; Cardona et al. 2014), albeit in different genes than those found in this study.

Second, in the Southeastern US, I see signals of selection on multiple genes that are directly and indirectly related to the human immune system, suggesting that pathogenic environments have been the strongest focus of adaptation in this population. The *A4GALT* gene, for example, codes for the P^k antigen which is part of the P blood group. This antigen is a receptor for shiga-like toxins produced by some strains of *Escherichia coli* as well as other verotoxins (Cooling 2015). The gene *IL1R1* codes for a cytokine receptor that plays a key role in the adaptive immune response. While not directly associated with the immune response, *VMP1* and *SLIT2* both have functions that are important during infections. The gene *VMP1* codes for a transmembrane protein that plays a key role in the process of autophagy, and this gene is known to be overexpressed during influenza infections of human cell lines (Hruz et al. 2008). The gene *SLIT2*, on the other hand, is involved in cell motility, including the movement of leukocytes during inflammatory responses caused by infection (Ye et al. 2010).

The suite of genes showing the strongest evidence of selection in the Southeastern US make sense given the colonial history of this region. The colonial period saw the introduction of a variety of diseases into the Americas, including smallpox, measles, influenza, pertussis, cholera, plague, typhus, yellow fever, diphtheria, malaria, and influenza (Ubelaker 2006). One recent spatial model of the colonial spread of epidemic diseases in North America (Jones 2014) suggests that such diseases were first introduced during European colonization to coastal areas of the Southeast in the early 16th century, spreading slowly towards the Appalachian mountains over the next 140 years and then moving very quickly across the interior Southeast. This model is consistent with the history of the region: after the initial Spanish colonization of the coastal Southeast in 1513, five documented expeditions (entradas) were undertaken to map the Southeast prior to 1545. Interaction between these entradas and Indigenous groups, along with the establishment of the Spanish mission system throughout the Southeast, likely contributed to the introduction and spread of multiple infectious diseases in this region. Several accounts of disease epidemics in the coastal Southeast are also described in historic records beginning in 1520 and continuing through the early 18th century (Stojanowski 2005). Groups in the interior Southeast likely avoided the first epidemics (Ramenofsky 1990). While the causal pathogens of many early epidemics remain unknown, accounts of some later epidemics allow the underlying cause to be identified, such as several accounts of a smallpox epidemic from the late 1690's that

report its spread from Virginia down into the Carolinas and across the Southeast into Mississippi (Kelton 2009). Altogether, historical documents, ethnohistoric records, Indigenous histories, and archaeological evidence demonstrate that these epidemics, in conjunction with other events and practices during the colonial era, contributed to a significant population decline and major sociopolitical changes in the region. By the 18th century, for example, many of the Indigenous groups interacting with the Spanish, including those forced into the mission system, had merged and the ethnogenesis of many of the modern Indigenous Southeastern groups was beginning (Galloway 1995). Our results suggest that the repeated epidemics may have also created significant selective pressures, influencing patterns of genetic variation at loci associated with the human immune response in these populations.

Pathogenic environments seem to have been a major selective force in Central Mexico as well, as many of the putatively selected genes in this population are also related to immune system pathways. In Mexico, the spread of European-introduced diseases began shortly after Spanish conquistador Hernán Cortés landed near the present-day city of Veracruz in 1519 and began his military campaign against the Aztec empire (Acuna-Soto et al. 2002). In Central Mexico, the first documented epidemic was smallpox in 1519-20, as Cortés marched towards the Aztec capital city of Tenochtitlan, located in present-day Mexico City in Central Mexico. This initial epidemic was followed by several subsequent smallpox

outbreaks, particularly in the late 16th and early 17th centuries (Acuna-Soto et al. 2002). After Tenochtitlan was conquered in 1521, it became known as Mexico City, capital of the Viceroyalty of New Spain, resulting in a large influx in people from both Europe and Africa (Restall and Schwaller 2011), no doubt bringing additional pathogens along with them.

This history likely contributed to genomic signatures of selection seen in the Central Mexico population in this study. Like in the Southeastern US, I see a strong signal of selection on the *IL1R1* gene (involved in the adaptive immune response) in Central Mexico. I also see a strong signal of selection on the mucin gene *MUC19* in the Central Mexico population. The GenomeRNAi database (Schmidt et al. 2013) shows that *MUC19* is associated with decreased vaccinia virus (VACV) infection. VACV is a close relative of the variola virus, the causal agent of smallpox, and recombinant versions of the VACV were used as a vaccine against smallpox until it was eradicated in the late 1970s (Jacobs et al. 2009). However, while these results are suggestive, I cannot be certain that smallpox was the selective agent for this sweep, because there were many more infectious diseases spreading through Mexico at this time (Acuna-Soto et al. 2002). Recent work using novel methods to search for ancient pathogen DNA in archaeological remains has been successful in finding some of these unknown pathogens, identifying the bacterium *Salmonella enterica* as a possible causal agent of the 16th Century 'cocoliztli' epidemic in southern Mexico, for example (Vågene et al. 2018). Future work may help us

identify the major pathogens afflicting the people of Central Mexico during colonial times.

Altogether, my analysis of genome-wide signals of selection in three indigenous North American populations found evidence for selection on genes related to cold and high latitude environments in the Alaska, but selection on genes related to immune function in the Southeastern US and Central Mexico. These results suggest that selective pressures have varied widely across the Americas, and further studies may find evidence of other adaptations in different environments on these continents. This study demonstrates the value of investigating selective pressures at a regional level in human populations.

Materials and Methods

Ethics and Community Engagement

This project was made possible through active and ongoing collaborations with members of the participating communities. Community authorities and representative bodies were consulted where appropriate, and all individual participants provided written informed consent for the types of analyses conducted in this study. The collection and analysis of samples from all communities was approved by the Institutional Review Board of the University of Texas at Austin (protocol #2012-05-0105). In some cases, exact sampling locations and/or community names are not reported to protect the privacy and anonymity of the communities and individuals participating in this research.

DNA Extraction and Genotyping

I extracted DNA from the saliva samples of 150 individuals from three populations in the Americas (Iñupiat from the North Slope of Alaska, N=35; Xaltocan in Central Mexico N=47; Southeastern US N=68) using the prepIT L2P kit (DNA Genotek), following the manufacturer's guidelines. Extracts were genotyped using the Affymetrix Human Origins Array.

Data QC and Admixture Analysis

SNPs not genotyped in the majority of study samples (--geno 0.1) were removed using PLINK v1.9 (Chang et al. 2015). For ADMIXTURE analysis, I further pruned the SNPS in high linkage disequilibrium (--indep-pairwise 50 1 0.1). A global ancestry analysis was conducted with ADMIXTURE 1.3.0 (Alexander et al. 2009) using the three study populations and West-Central African (Yoruba) and Western European (Spanish, French) outgroups, drawn from the previously published Affymetrix Human Origins Array dataset to represent the likely sources of recent admixture in the study populations. I restricted subsequent analyses to individuals with <20% cluster membership in both the European and African components based on my K=3 ADMIXTURE analysis to maximize sample sizes and to minimize the effects of admixture on my inferences of selection. After restricting the dataset based on these criteria, my sample sizes were N=27 for the Alaskan Arctic (Iñupiat) population, N=43 for the Central Mexico (Xaltocan) population, and N=47 for the Southeastern US population.

Haplotype Phasing and Selection Analyses

I phased this dataset using SHAPEIT2 (Delaneau et al. 2012) with the default parameters. The phased data were then annotated with ancestral allele information using aa_annotate.py (Cadzow et al. 2014). iHS values were calculated for each population using hapbin (Maclean et al. 2015). F_{ST} values were calculated for each of the three sampled populations with an ingroup (33 Peruvian individuals without any evidence of recent European or African genetic ancestry, selected from the 1000 Genomes Project dataset) and with an outgroup (50 Han Chinese individuals from the 1000 Genomes Project) using vcftools (Danecek et al. 2011). The Population Branch Statistic (Yi et al. 2010) was then calculated for the three sampled Native North American populations.

The top 1% of PBS and iHS scores were identified for each population respectively, and then cross referenced so that further analysis was done only for SNPs in the top 1% of scores in both tests for selection. This cross-referencing should reduce my chances of reporting selection results that are actually false positives since each statistic has different underlying assumptions.

Gene Annotation and Pathway Enrichment Analysis

I used the Ensembl Variant Effect Predictor (McLaren et al. 2016) to assign gene names to SNPs in the top 1% of PBS and iHS scores. In order to identify any metabolic pathways that are overrepresented in the putatively selected genes in each of the three study populations, I performed a path enrichment analysis using

the WebGestalt platform, with the parameters: hsapiens > overrepresentation enrichment Analysis > geneology > Biological process and hsapiens > overrepresentation enrichment Analysis > Pathway > KEGG.

References

1000 Genomes Project Consortium, et al. (2015) A global reference for human genetic variation. Nature 526(7571):68-74.

Acuna-Soto R, Stahle DW, Cleaveland MK, Therrell MD (2002) Megadrought and megadeath in 16th century Mexico. Emerg Infect Dis 8(4):360-362.

Ahn J, et al. (2010) Genome-wide association study of circulating vitamin D levels. Hum Mol Genet 19:2739–2745.

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19(9):1655-1664.

Anderson CNK, Ramakrishan U, Chan YL, Hadly EA. (2005) Serial SimCoal: a population genetic model for data from multiple populations and points in time. Bioinform 21:1733-1734.

Bierhorst J. (1992) Annals of Cuauhtitlan. History and mythology of the Aztecs: The Codex Chimalpopoca. The University of Arizona Press, Tucson, AZ.

Bigham A, et al. (2010) Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. PLoS Genet 6(9):e1001116.

Bolnick DA, et al. 2012. Non-destructive sampling of human skeletal remains yields ancient nuclear and mitochondrial DNA. Am J Phys Anthropol 147:293-300.

Bolnick DA, Raff JA, Springs LC, Reynolds AW, Miró-Herrans AT (2016) Native American genomics and population histories. Ann Rev Anthropol 45:319-340.

Brumfiel EM (2005a) Introduction: production and power at Postclassic Xaltocan. Production and power at Postclassic Xaltocan, ed Brumfiel EM (Instituto Nacional de Antropología e Historia, Mexico City, Mexico), pp 27-41.

Brumfiel EM (2005b) Conclusions: production and power at Xaltocan. Production and power at Postclassic Xaltocan, ed Brumfiel EM (Instituto Nacional de Antropología e Historia, Mexico City, Mexico), pp 349-367.

Brumfiel EM (2005c) Opting in and opting out: Tula, Cholula, and Xaltocan. Settlement, subsistence, and social complexity: essays honoring the legacy of Jeffrey R. Parsons, ed Blanton RE (Cotsen Institute of Archaeology at UCLA, Los Angeles, CA), pp 63-88.

Brumfiel EM. (2007) Otras investigaciones en Xaltocan, 1997-2005. Estrategias de las unidades domésticas en Xaltocan postclásico, México: informe final al Instituto Nacional de Antropología e Historia, ed Brumfiel EM, (Instituto Nacional de Antropología e Historia, Mexico City), pp 71-109.

Brumfiel EM, Rodríguez-Alegría E. (2010) Estrategias de las élites y cambios políticos en Xaltocan, México. Field report on file at the Consejo de Arqueología, Instituto Nacional de Antropología e Historia, Mexico.

Cadzow M, et al. (2014) A bioinformatics workflow for detecting signatures of selection in genomic data. Front Genet 5:293.

Cardona A, et al. (2014) Genome-wide analysis of cold adaptation in indigenous Siberian populations. PloS One 9(5):e98076.

Carrasco-Pizana P. (1987) Los otomíes: cultura e historia prehispánica de los pueblos mesoamericanos de habla otomiana. Toluca, Mexico: Ediciones del Gobierno del Estado de México.

Chang CC, et al. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4:7.

Cheng JB, Levine MA, Bell NH, Mangelsdorf DJ, Russell DW (2004) Genetic evidence that the human CYP2R1 enzyme is a key vitamin D 25-hydroxylase. Proc Natl Acad Sci USA 101(20):7711-7715.

Cooling L (2015) Blood groups in infection and host susceptibility. Clin Microbiol Rev 28(3):801-870.

Curcio-Nagy LA. (2011) The kingdom of New Spain in the seventeenth century. A companion to Mexican history and culture, ed Beezley WH (Wiley-Blackwell, Malden, MA), pp 209-229.

Danecek P, et al. (2011) The variant call format and VCFtools. Bioinformatics 27(15):2156-2158.

De Lorenzo F, Kadziola Z, Mukherjee M, Saba N, Kakkar VV (1999) Haemodynamic responses and changes of haemostatic risk factors in cold-adapted humans. QJM 92(9):509-513.

De Lucia K. (2010) A child's house: social memory, identity, and the construction of childhood in Early Postclassic Mexican households. Am Anthropol 112:607-624.

De Lucia K, Overholtzer L. (2014) Everyday action and the rise and decline of ancient polities: household strategy and political change in Postclassic Xaltocan, Mexico. Anc Mesoam 25:441-458.

Delaneau O, Marchini J, Zagury JF (2012) A linear complexity phasing method for thousands of genomes. Nat Methods 9(2):179-81.

Endfield GH. (2008) Climate and society in colonial Mexico. Blackwell Publishing, Malden, MA.

Excoffier L, Novembre J, Schneider S. (2000) SIMCOAL: a general coalescent program for simulation of molecular data in interconnected populations with arbitrary demography. J Heredity 91:506-509.

Excoffier L, Lischer HEL. (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetic analyses under Linux and Windows. Mol Ecol Res 10:564-567.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. (2013) Robust demographic inference from genomic and SNP data. PLoS Genet 9(10):e1003905.

Fan S, Hansen MEB, Lo Y, Tishkoff SA (2016) Going global by adapting local: A review of recent human adaptation. Science 354(6308):54-59.

Field Y, et al. (2016) Detection of human adaptation during the past 2000 years. Science 354(6313):760-764.

Fu Q, et al. (2013) DNA analysis of an early modern human from Tianyuan Cave, China. Proc Natl Acad Sci USA 110(6):2223-2227.

Fumagalli M, et al. (2015) Greenlandic Inuit show genetic signatures of diet and climate adaptation. Science 349(6254):1343-1347.

Galloway P (1995) Choctaw Genesis, 1500-1700. University of Nebraska Press, Lincoln, NE.

García-Martínez B. 2010. Los años de la conquista. Nueva historia general de México, ed Velásquez-García E (El Colegio de México, Mexico City, Mexico), pp 169-215.

Gibson C. (1964) The Aztecs under Spanish rule: a history of the Indians of the Valley of Mexico 1519-1810. Stanford University Press, Stanford, CA.

Gusev A, et al. (2009) Whole population, genome-wide mapping of hidden relatedness. Genome Res 19(2):318-326.

Haak W, et al. (2015) Massive migration from the steppe was a source for Indo-European languages in Europe. Nature 522:207-211.

Hancock AM, et al. (2011) Adaptations to climate-mediated selective pressures in humans. PLoS Genet 7(4):e1001375.

Hansen et al. (2017) The indigenous world 2017, (The International Work Group for Indigenous Affairs; Copenhagen, Denmark), pp 116-126.

Hellenthal G, et al. (2014) A genetic atlas of human admixture history. Science 343(6172):747–751.

Hicks F. (1994) Xaltocan under Mexican domination, 1435-1520. Caciques and their people, ed Marcus J, Zeitlin JF (Museum of Anthropology, University of Michigan, Ann Arbor, MI), pp 67-85.

Hicks F. (2005) Mexico, Acolhuacan, and the rulership of Late Postclassic Xaltocan. Production and power at Postclassic Xaltocan, ed Brumfiel EM (Instituto Nacional de Antropología e Historia, Mexico City, Mexico), pp 195-206.

Hruz T, et al. (2008) Genevestigator V3: a reference expression database for the meta-analysis of transcriptomes. Adv Bioinformatics 420747. (https://genevisible.com/perturbations/HS/Gene%20Symbol/VMP1; accessed 18/08/2017)

Ixtlilxóchitl ACF. (1975) Obras históricas, tomo 1, translated by O'Gorman E. Universidad Autónoma de México, Mexico City, Mexico.

Ixtlilxóchitl ACF. (1977) Obras históricas, tomo 2, translated by O'Gorman E. Universidad Autónoma de México, Mexico City, Mexico.

Jablonski NG, Chaplin G (2010) Human skin pigmentation as an adaptation to UV radiation. Proc Natl Acad Sci USA 11(107):8962-8968.

Jacobs BL, et al. (2009) Vaccinia virus vaccines: past, present and future. Antiviral Res 84(1):1-13.

Jiang X, et al. (2018) Genome-wide association study in 79,366 European-ancestry individuals informs the genetic architecture of 25-hydroxyvitamin D levels. Nat Commun 9:260.

Jones EE (2014) A spatiotemporal analysis of old world diseases in North America, A.D. 1519-1807. Am Antiq 3:487-506.

Kalinowski ST, Taper ML, Marshall TC. (2007) Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. Mol Ecol 16:1099-1106.

Karlsson EK, et al. (2013) Natural Selection in a Bangladeshi population from the cholera-endemic Ganges River delta. Sci Transl Med 5(192): 192ra86.

Kelton P (2009) Shattered and infected: epidemics and the origins of the Yamasee War, 1696–1715. Mapping the Mississippian Shatter Zone: The Colonial Indian Slave Trade and Regional Instability in the American South, eds Ethridge R, Shuck-Hall SM(University of Nebraska Press, Lincoln), pp 312–332.

Kornelissen TS, Albrechtsen A, Nielsen R. (2014) ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinform 15:356.

Konovalov DA, Manning C, Henshaw MT. (2004) KINGROUP: a program for pedigree relationship reconstruction and kin group assignments using genetic markers. Mol Ecol Notes 4:779-782.

Li H, Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinform 25(14):1754-1760.

Lindo J, et al. (2016) A time transect of exomes from a Native American population before and after European contact. Nat Commun 7:13175.

Lockhart J. (1992) The Nahuas after the conquest. Stanford University Press, Stanford, CA.

Loh P-R, et al. (2013) Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium. Genetics 193(4):1233-1254.

Maclean CA, Chue Hong NP, Prendergast JG (2015) hapbin: An efficient program for performing haplotype-based scans for positive selection in large genomic datasets. Mol Biol Evol 32(11):3027-3029.

Maples BK, Gravel S, Kenny EE, Bustamante CD. (2013) RFMix: A discriminative modeling approach for rapid and robust local ancestry inference. Am J Hum Gen 93(2):278-288.

Mata-Míguez J, Overholtzer L, Rodríguez-Alegría E, Kemp BM, Bolnick DA. (2012) The genetic impact of Aztec imperialism: ancient mitochondrial DNA evidence from Xaltocan, Mexico. Am J Phys Anthropol 149:504-516.

Mata-Míguez J. (2016) Assessing the demographic and genetic impact of state expansion in pre-Hispanic and colonial Mexico. PhD Dissertation, University of Texas at Austin, Austin, TX.

McLaren W, et al. (2016) The ensembl variant effect predictor. Genome Biol 17(1):122.

Morehart CT, Eisenberg DTA (2010) Prosperity, power, and change: modeling maize at Postclassic Xaltocan, Mexico. J Anthropol Arch 29:94-112.

Moreno-Estrada A, et al. (2014) The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. Science 344(6189):1280-1285.

Mychaleckyj JC, et al. (2017) Genome-wide analysis in Brazilians reveals highly differentiated Native American genome regions. Mol Biol Evol 34(3):559-574.

Orozco L, et al. (in prep.) Genomic data from Mexico.

Overholtzer L. (2012) Empire and everyday material practices: a household archaeology of Aztec and Spanish imperialism at Xaltocan, Mexico. PhD Dissertation, Northwestern University, Chicago, IL.

Overholtzer L. (2013) Archaeological interpretation and the rewriting of history: deimperializing and decolonizing the past at Xaltocan, Mexico. Am Anthropol 115:481-495.

Overholtzer L. (2015) Agency, practice, and chronological context: a Bayesian approach to household chronologies. J Anthropol Arch 37:37-47.

Patterson N, Price AL, Reich D. (2006) Population structure and eigenanalysis. PLoS Genet 2(12):e190.

Patterson N, et al. (2012) Ancient admixture in human history. Genetics 192(3):1065-1093.

Perry GH, et al. (2014) Adaptive, convergent origins of the pygmy phenotype in African rainforest hunter-gatherers. Proc Natl Acad Sci USA 111(35):3596-3603.

Peter BM, Petkova D, Novembre J. (2018) Genetic landscapes reveal how human genetic diversity aligns with geography. bioRxiv (https://doi.org/10.1101/233486)

Petkova D, Novembre J, Stephens M. (2016) Visualizing spatial population structure with estimated effective migration surfaces. Nat Gen 48:94-100.

Pickrell JK, et al. (2009) Signals of recent positive selection in a worldwide sample of human populations. Genome Res 19(5):826-837.

Pickrell JK, Pritchard JK. (2012) Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. PLoS Genet 8(11): e1002967.

Raghavan M, et al. (2015) Genomic evidence for the Pleistocene and recent population history of Native Americans. Science 349(6250).

Ramenofsky AF (1990) Loss of innocence: explanations of differential persistence in the sixteenth-century Southeast. Columbian Consequences: The Spanish Borderlands in Pan-American Perspective, Vol. 2, ed Thomas DH (Smithsonian Institution Press, Washington, D.C.), pp 31–48.

Rasmussen M, et al. (2014) The genome of a late Pleistocene human from a Clovis burial site in western Montana. Nature 506(7487): 225-229.

Rasmussen M, et al. (2015) The ancestry and affiliations of Kennewick Man. Nature 523:455-458.

Reich D, et al. (2012) Reconstructing native American population history. Nature 488:370–374.

Restall M, Schwaller R. (2011) The gods return: conquest and conquest society (1502- 1610). A companion to Mexican history and culture, ed Beezley WH (Wiley-Blackwell, Malden, MA), p 195-208.

Roland N, Hofreiter M. (2007) Ancient DNA extraction from bones and teeth. Nat Protocol 2:1756-1762.

Rohland N, Harney E, Mallick S, Nordenfelt S, Reich D. (2015) Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. Philos Trans R Soc Lond B Biol Sci. 370(1660).

Sanders WT, Parsons JR, Santley RS. (1979) Basin of Mexico: ecological processes in the evolution of a civilization. Academic Press, New York, NY.

Sawyer S, Krause J, Guschanski K, Savolainen V, Pääbo S. (2012) Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. PLoS One 7(3):e34131.

Schiffels S, et al. (2016) Iron Age and Anglo-Saxon genomes from East England reveal British migration history. Nat Commun 7:10408.

Schlebusch CM, et al. (2015) Human adaptation to arsenic-rich environments. Mol Biol Evol 32(6):1544-1555.

Schmidt EE, et al. (2013) GenomeRNAi: a database for cell-based and in vivo RNAi phenotypes, 2013 update. Nucleic Acids Res 41:D1021-6. (http://www.genomernai.org/v16/genedetails/2/283463 ; accessed 18/08/2017)

Skoglund P, et al. (2015) Genetic evidence for two founding populations of the Americas. Nature 525:104-108.

Skoglund P, Matheison M. (2018) Ancient genomics: a new view into human prehistory and evolution, Ann Rev Genom Hum Genet.

Socolow SM. 1996. Introduction to the rural past. The countryside in colonial Latin America, ed Hoberman LS, Socolow SM (University of New Mexico Press, Albuquerque, NM), pp 3-18.

Stelzer G, et al. (2016) The GeneCards suite: from gene data mining to disease genome sequence analyses. Curr Protoc Bioinformatics 54:1.30.1-1.30.33. (www.genecards.com)

Stevens TH, Forgac M. (1997) Structure, function and regulation of the vacuolar ATPase. Ann Rev Cell Devol Biol 13:779-808.

Stojanowski CM (2005) Biocultural Histories in La Florida: A Bioarchaeological Perspective. (University of Alabama Press, Tuscaloosa).

Ubelaker DH (2006) Population size, contact to nadir. Handbook of North American Indians, Volume 3: Environment, Origins, and Population: Environment, Origins, and Population, ed Ubelaker DH (Smithsonian Institution Press, Washington, D.C.), pp 694-791.

Vågene AJ, et al. (2018) Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico. Nature Ecol Evo 2(1):1-9.

Voight CF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4(3):e72.

Wang J, Vasaikar S, Shi Z, Greer M, Zhang B, (2017) WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. Nucleic Acids Res 45(1):130-137.

Wang TJ, et al. (2010) Common genetic determinants of vitamin D insufficiency: a genome-wide association study. Lancet 376:180–188.

Ye BQ, Geng ZH, Ma L, Geng JG (2010) Slit2 regulates attractive eosinophil and repulsive neutrophil chemotaxis through differential srGAP1 expression during lung inflammation. J Immunol 185(10):6394-6405.

Yi X, et al. (2010) Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329(5987):75-78.

Zhang YY, et al. (2014) BKCa and hEag1 channels regulate cell proliferation and differentiation in human bone marrow-derived mesenchymal stem cells. J Cell Physiol 229(2):202-212.